

A HUMÁN TRANZMEMBRÁN FEHÉRJE SZERKEZETI TÉR VIZSGÁLATA A TMALPHA FOLD ADATBÁZIS FELHASZNÁLÁSÁVAL

36. OTDK Biológia Szekció

Jambrich Márton András

Budapesti Műszaki és Gazdaságtudományi Egyetem, Vegyészmérnöki és Biomérnöki Kar

Biomérnök MSc, I. évf.

Témavezető: Dr. Dobson László

Eötvös Loránd Kutatási Hálózat, Természettudományi Kutatóközpont

Enzimológiai Intézet, Fehérje Bioinformatika Kutatócsoport



Budapest

2023

Tartalomjegyzék

1. Absztrakt	2
2. Abstract	3
3. Bevezetés	4
3.1. A feltekeredési (folding) probléma.....	4
3.2. Transzmembrán fehérjék	6
3.3. Fehérjék háromdimenziós szerkezetének kísérletes meghatározása	7
3.4. Fehérjék háromdimenziós szerkezetének számítógépes meghatározása.....	8
3.5. A szerkezeti tér lefedettsége az AlphaFold2 előtt	9
3.6. Az AlphaFold2 szerkezetbecslő eljárás	10
4. Célkitűzés.....	12
5. Adatok és módszerek	13
5.1. Bioinformatikai módszerek	13
5.2. Adatbázisok	14
5.2.1. Széles körben használt adatbázisok.....	14
5.2.2. Specifikus adatbázisok	14
5.2.2.1. PDBTM.....	14
5.2.2.2. Orientations of Proteins in Membranes (OPM).....	15
5.2.2.3. Human Transmembrane Proteome (HTP).....	15
5.2.2.4. AlphaFold Database (AFDB)	15
5.2.2.5. Transzmembrán AlphaFold adatbázis (TmAlphaFold database).....	15
5.3. A kiértékeléshez használt statisztikai módszerek, metrikák.....	17
5.4. Adatszettek	18
5.4.1. AlphaFold2 humán TM fehérjéken való kiértékelése (humán adatszett).....	18
5.4.2. Homológ fehérjék keresése AlphaFold2 szerkezetek segítségével (CATH adatszett)	18
6. Eredmények és kiértékelés.....	19
6.1. A TmAlphaFold adatbázis kiértékelése humán transzmembrán fehérjéken	19
6.2. A humán TM szerkezeti tér lefedettségének minősége az AF2 előtt és után	21
6.3. Szekvencia és szerkezet alapú módszerek összehasonlítása homológ fehérjék keresésére.....	25
6.4. Gyengébb minőségű szerkezetbecslések elemzése szerkezeti, funkcionális és gyógyszertervezési szempontból	30
7. Összefoglalás	35
8. Köszönetnyilvánítás	36
9. Hivatkozások.....	37

1. Absztrakt

A feltekeredés (folding) probléma a biológia több mint 60 éves megválaszolatlan kérdése, ami arra keresi a választ, hogyan határozza meg egy fehérje aminosav szekvenciája annak térszerkezetét. A kísérletes technikák fejlődésével a megoldott fehérjeszerkezetek száma exponenciálisan nő, azonban a feltekeredés biofizikai leírására nem létezik olyan modell, ami belátható időn belül képes meghatározni akár csak egyetlen fehérje szerkezetét. A DeepMind 2020-ben kifejlesztett mély gépi tanuláson alapuló módszere, az AlphaFold2 (AF2) hatalmas mértékben javította a szerkezetbecslés pontosságát, és már egy évvel a megjelenése után is megreformálni látszik a szerkezeti biológia területét.

Mind a kísérletes, mind a számítási módszerek esetén szűk keresztmetszetet jelent a transzmembrán (TM) fehérjék szerkezetének meghatározása. Ezen fehérjék peptidlánca két különböző környezetben halad át, vízdoldható és lipoldható környezetben. Ezen körülmények között a feltekeredés leírásához speciális biofizikai megfontolásokat kell alkalmazni, amely figyelembe veszi az eltérő fázisokat, továbbá ezen fehérjék kísérletes szerkezetmeghatározása is speciális körülményeket kíván meg. Bár a TM fehérjék a teljes humán proteom 25-30%-át teszik ki, továbbá ezek a leggyakoribb gyógyszer-célpontok, mégis mindössze a megoldott szerkezetek ~4 százaléka tartozik ebbe a csoportba, és a számítógépes becslések pontossága is elmarad ezen fehérjék esetén a globuláris fehérjékhez képest.

Bár az AF2 minden fehérjének képes modellezni a szerkezetét, de mint minden módszer, ez is hibával terhelt lehet. TM fehérjék esetében a hibák egyik fő forrása lehet, hogy a tanítás során az ismert TM szerkezetek száma nagyon alacsony volt. A nemrég fejlesztett TmAlphaFold adatbázis tartalmazza az AF2 által becsült TM fehérje szerkezetek membrán lokalizációját, valamint listázza azokat a szerkezeti hibákat, amelyek téves becslésre utalnak. Ezek az adatok lehetővé teszik az AF2 által becsült TM fehérje szerkezetek proteom szintű analízisét.

Munkám során bemutatom, hogyan változtatta meg az AF2 a humán TM fehérjék szerkezeti terének lefedettségét, továbbá megmutatom, melyek azok a TM fehérjék és fehérjecsaládok, amelyeknél az AF2 pontossága alacsonyabb és nem rendelkeznek kísérletesen megoldott szerkezettel, ugyanis ezen fehérjék szerkezetének megoldásából profitálhat a legtöbbet a szerkezeti biológia és a gyógyszerhatóanyag fejlesztés.

2. Abstract

The folding problem has been an open question in biology for more than 60 years, looking to answer how the amino acid sequence of a protein determines its structure. With the advancement of experimental techniques, the number of solved protein structures is increasing exponentially, yet there is no model for the biophysical description of folding that is capable of determining the structure of even a single protein in a reasonable amount of time. DeepMind's deep machine learning based method AlphaFold2 (AF2), developed in 2020, has vastly improved the accuracy of structure prediction and, just one year after its release, looks set to reform the field of structural biology.

A bottleneck for both experimental and computational methods is the determination of the structures of transmembrane (TM) proteins. The peptide chains of these proteins pass through two different environments: water-soluble and lipid-soluble. To describe the folding under these conditions, special biophysical considerations must be applied that take into account the different phases, furthermore the experimental structure determination of these proteins also requires special conditions. Although TM proteins account for 25-30% of the total human proteome, and are the most common drug targets, only ~4% of the solved structures fall into this group, and the accuracy of computational predictions for these proteins is also below that of globular proteins.

Although AF2 can model the structure of all proteins, like all methods, it can be subject to error. One of the main sources of possible errors in case of TM proteins may be that the number of known TM structures was very low during the training. The recently developed TmAlphaFold database contains the membrane localizations of TM protein structures estimated by AF2, as well as a list of structural errors that indicate incorrect prediction. This data allows for proteome-wide analysis of AF2-estimated TM protein structures.

Throughout my work, I will illustrate how AF2 has changed the coverage of the structural space of human TM proteins and gather proteins and protein families with lower AF2 accuracy and no experimentally solved structures, since the resolution of the structures of these proteins could be the most beneficial for structural biology and drug development.

3. Bevezetés

3.1. A feltekeredési (folding) probléma

Az újgenerációs szekvenálás (next generation sequencing: NGS) megjelenésének hatására ma már rutin feladat a különböző genomok feltárása. A centrális dogma értelmében az így kapott nukleotid szekvenciákból könnyedén megtudhatjuk a genomban kódolt fehérjék aminosavsorrendjét, ami az ismert szekvenciájú fehérjék mennyiségének exponenciális növekedését idézte elő.

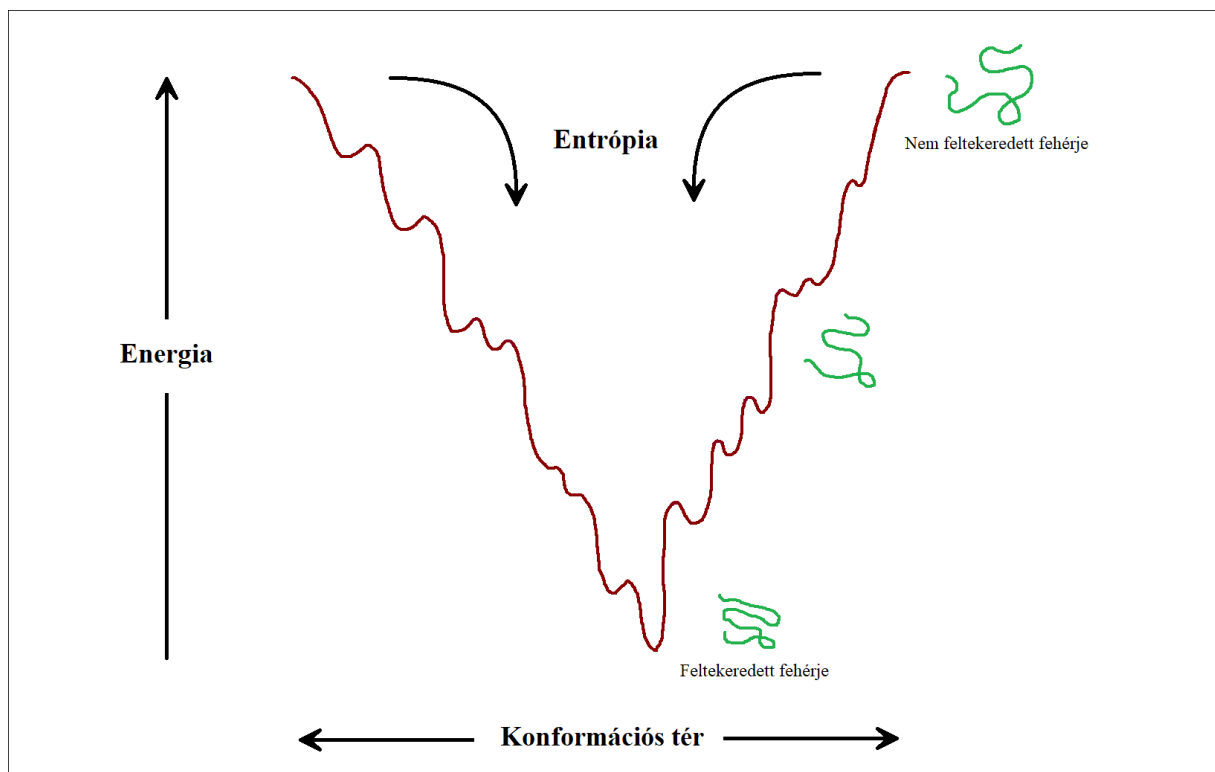
A fehérjék lineáris polimerek, melyeket 20 különböző aminosav épít fel. Háromdimenziós szerkezetük jellemzése szempontjából megkülönböztetünk elsődleges, másodlagos, harmadlagos és negyedleges fehérjeszerkezeti szinteket. Az elsődleges szerkezet a fehérje peptid kötéssel kapcsolódó aminosavjainak sorrendjét írja le. A másodlagos szerkezet kisebb, lokális konformációkat jellemez a fehérjeláncon, ezeket hidrogén hidak tartják össze. Ezek a szegmensek három torziós szöggel írhatóak le, amelyek közül egy állandó és kettő változik, így létrehozva periodikusan ismétlődő szakaszokat, az alfa-hélixet és a béta-lemezt. A harmadlagos szerkezet a másodlagos szerkezetek kapcsolódására utal és a fehérje háromdimenziós szerkezetét adja meg, és abban az összes atom helyzetét. A harmadlagos szerkezet fenntartásában a domináns erő a hidrofób kölcsönhatás. A negyedleges szerkezetet a több alegységes fehérjékre értelmezzük, az alegységek egymáshoz képesti elhelyezkedését jellemzi. Vannak egyéb szerkezeti szintek is, amelyek nem sorolhatóak be a fent említett rendszerbe, ilyenek például a motívumok, amik a másodlagos szerkezetek kapcsolódásának sorrendjét jellemzik (pl.: alfa-alfa, alfa-beta-alfa stb.), valamint a domének, amelyek önálló feltekeredésre képes funkcionális egységek.

Azt a feltevést, hogy a fehérjék harmadlagos szerkezetét az elsődleges szerkezet kódolná, Christian Anfinsen bizonyította be kísérletes úton. A ribonukleáz-A fehérjét vizsgálta oly módon, hogy ureát adott hozzá, ami felszakította a fehérjében lévő nemkovelens kötéseket, majd béta-merkaptóetanollal redukálta a fehérje diszulfid-hídjait is, így kitekeredett a fehérje és elvesztette enzim funkcióját. Az urea és a béta-merkaptóetanol dialízissel való eltávolítása után az enzim visszanyerte a funkcióját, tehát az elsődleges szerkezet valóban magában hordozta a feltekeredéshez szükséges információt [1]. A Levinthal paradoxon egy gondolatkísérlet, egy 101 aminosavas fehérje feltekerése, ebben 100 peptidkötés található. Tegyük fel, hogy minden kötés csak 2 féle állapotba rendeződhet, és egy konformáció kipróbálása 1 ns-ot vesz igénybe [2]. Az összes eset kipróbálása így $4 \cdot 10^{13}$ évig tartana. Ez

alapján beláthatjuk, hogy egy fehérje natív szerkezete véletlenszerű kereséssel nem található meg, ez irányított folyamat kell, hogy legyen. A feltekeredési (folding) probléma tehát annak a kérdése, hogy pontosan hogyan lehet leírni a fehérjék feltekeredését termodinamikai vagy kinetikai szempontból, illetve, hogy ez becsülhető-e számítógépes módszerrel.

Fontos megemlíteni, hogy bizonyos fehérjék nem írhatóak le tökéletesen a fenti szabályokkal és megközelítésekkel. Ezeket a dinamikusan változó, flexibilis fehérjéket (vagy szakaszokat) rendezetlen fehérjéknek (vagy régióknak) nevezzük: nem rendelkeznek időben állandó, harmadlagos szerkezettel, több konformáció közötti folyamatos átalakulásban vannak, és éppen így látják el a funkciójukat.

A folding reprezentálható egy tölcsérként, amiben vízszintesen (x-tengely) a konformációs koordináták találhatók, függőlegesen (y-tengely) pedig a szabadentalpia (1. ábra). A nem feltekeredett fehérjére jellemző a magas energiaállapot és széles konformáció tartomány. A feltekeredett fehérje a tölcsér alján van, és csak kevés optimális szerkezetet vehet fel (a rendezetlen fehérjéknek pedig energetikai szempontból több optimális állapota is van) [3].



1. ábra. A folding tölcsér modellje

3.2. Transzmembrán fehérjék

Az élő sejteket és a bennük található egyes organelleket membrán határolja, ami egy, a vízdíható anyagok számára gyakorlatilag átjárhatatlan réteg. A membránok főként foszfolipidekből állnak, melyek amfipatikus tulajdonságaikból adódóan képesek kettős lipid réteget létrehozni. Emellett a membránokban megtalálunk még egyéb lipideket, illetve szénhidrátokat és fehérjéket is. A membránnak fontos szerepe van az egysejtű és többsejtű fajokban is, elhatárolja a sejtet, vagy sejtorganelleket a külső tértől, illetve kontrollált információ transzportot (anyagtranszportot, jelátvitelt, energiatermelést) biztosít külső, és belső oldala között. Többek között az ilyen feladatok ellátására specializálódtak a membránban elhelyezkedő transzmembrán (TM) fehérjék, amelyek ennél fogva igen gyakori gyógyszer-célpontok, a gyógyszerek több mint fele TM fehérjéken keresztül fejtik ki hatását [4].

A TM fehérjék polipeptid lánc egyszer, vagy többször teljesen átszeli a membrán kettős lipid réteget, ami a membránban elhelyezkedő, töltött funkciós csoporttal rendelkező aminosavak számára kedvezőtlen, így a membránban elhelyezkedő részek alfa-hélix vagy béta-szál másodlagos szerkezetet vesznek fel, a membránra nagyjából merőleges lefutással. A membrán szélessége nagyjából 30 Å, ezt alfa-hélix másodlagos szerkezet mellett jellemzően 20-25 aminosavval lehet átérni, béta-szál esetén 8-10 aminosavval. Léteznek azonban kivételek, elképzelhető például extrém módon megdől alfa-hélix, ahol sokkal hosszabb membrán régióra számíthatunk. Vannak speciális elemek is, ahol nem minden membránnal érintkező rész halad keresztül a membránon, ilyenek például a visszahurkoló régiók, amik a membrán azonos oldalán mennek be, majd lépnek ki onnan, illetve a határfelületi hélixek, melyek a membrán határsíkjával párhuzamos amfipatikus hélixek, ezek gyakran lipid-fehérje kölcsönhatásokban játszanak szerepet. A humán TM fehérjék többsége kizárólag alfa-hélix másodlagos szerkezeti elemekkel rendelkezik, a béta-hordó szerkezetű TM fehérje csak a mitokondrium belső membránjában fordul elő. A továbbiakban, amennyiben TM fehérjéről beszélek, azzal mindig az alfa-helikális TM fehérjékre utalok.

3.3. Fehérjék háromdimenziós szerkezetének kísérletes meghatározása

A fehérjék szerkezetének kísérletes megoldására többféle módszer létezik, ezeknek célja a szerkezetben lévő atomok koordinátáinak minél pontosabb meghatározása. Az egyik ilyen technika a röntgen krisztallográfia, melynek során a vizsgált fehérjét kristályrácsban rögzítjük, majd röntgensugárzásnak tesszük ki. A fehérje kristályrácsán a röntgensugár szóródik, egy egyedi diffrakciós mintázatot hozva létre, melyből felépíthetjük a fehérje elektronsűrűség térképét. Ennek a térképnek a pontjaiba illeszthetők be a fehérjében szereplő atomok, így meghatározva az eredeti háromdimenziós szerkezetet. A módszernek egyik hátránya, hogy ez a fehérje kristályosítása miatt csak egy statikus képet tud nyújtani a fehérje szerkezetéről, illetve, hogy TM fehérjék esetén a vizsgálat előtt egy detergenssel el kell távolítani a membránt, és ez befolyásolja a fehérje natív szerkezetét.

Egy másik módszer a mágneses magrezonancia (NMR) vizsgálat, ami azt az elvet használja ki, hogy ha a protonok vagy neutronok száma egy atomban páratlan, akkor ezek mágneses térbe helyezve információt adnak a szomszédos kapcsolataikról. Leegyszerűsítve ez olyan módon történik, hogy mágneses térben gerjesztődnek, árnyékolás jön létre, és az így kapott értéket lehet egy referencia (árnyékolás nélküli) értékhez hasonlítani, majd az így kapott kémiai eltolódás értékekkel egy megszorítás halmaz állítható fel az egyes atomok távolságára. Ezeket a távolságmegszorításokat felhasználva kaphatjuk meg a molekula lehetséges szerkezeteit, ebből kifolyólag a fehérje dinamikus mozgását is látjuk, ami az NMR egyik legnagyobb előnye. A módszer hátránya, hogy NMR-rel csak kisebb molekulák vizsgálhatóak, a TM fehérjék pedig gyakran túl nagyok az ilyen típusú szerkezetmeghatározáshoz.

Krio-elektronmikroszkópia (krio-EM) esetén először olyan gyorsan lefagyasztjuk a fehérje molekulát, hogy a benne lévő víz ne tudjon kristályosodni, így vitrifikált amorf víz képződik. Röntgen sugár helyett ilyenkor elektronokkal bombázzuk a fagyasztott fehérje molekulát különböző irányokból, majd az így kapott kétdimenziós vetületekből alkotjuk meg a háromdimenziós modellt. Ez az eljárás az utóbbi években jelentős előrelépést hozott TM fehérjék szerkezetmeghatározásában, ugyanis ilyen módon a natívhoz közeli állapotban tudjuk megfigyelni a fehérjét. Hátránya, hogy jelentősen gyengébb a felbontása az egyéb szerkezetmegoldási módszerekhez képest.

3.4. Fehérjék háromdimenziós szerkezetének számítógépes meghatározása

A fehérjék háromdimenziós szerkezetének számítógépes megoldása egy rendkívül komplex, nagyon nagy számítási igényű feladat, ezért egyedi megoldások szükségesek ahhoz, hogy ez lehetségessé váljon. A klasszikus szerkezetmodellező módszerek három kategóriába sorolhatóak: homológia modellezés, fold felismerés és *de novo* szerkezetbecslés.

A homológia modellezés során azt az elvet használjuk ki, hogy a szerkezet erősebben konzervált, mint a szekvencia, tehát még alacsony szekvenciaazonosságok mellett is megfigyelhetők nagyon hasonló szerkezetek. Ennek megfelelően a modellezni kívánt szekvenciánkhoz igyekszünk olyan szekvenciát keresni, ami legalább 20-40%-os azonosságot mutat, és rendelhető hozzá megoldott szerkezet. Ennek a szerkezetnek a segítségével próbáljuk modellezni a fehérjénket, bár ezzel a módszerrel várhatóan lesznek olyan szakaszok/hurkok, amelyek hiányoznak majd a modellszerkezetből, amiket majd pótolni kell.

Amennyiben homológia modellezéssel nem tudjuk modellezni a szerkezetünket, a fold felismerés módszert is alkalmazhatjuk. Ilyenkor egy reprezentatív szerkezeti készletet használunk, amelynek összes szerkezetéhez hozzápróbáljuk szekvenciánkat. A szekvenciánkat végigfűzzük ('threading') az egyes szerkezeteken, és egy energiafüggvénnyel leírjuk az összes szerkezet összes pozíciójában, hogy ott mennyi lenne a fehérje energiája. Ezek közül a legkisebb energiaállapot tartozhat a natív állapothoz. A módszer során mindig lesz egy ilyen legkisebb energiaértékű állapot, de egyáltalán nem biztos, hogy ez valóban helyes szerkezetnek felel meg. Ennek érdekében feltételeznünk kell egy bizonyos energiaérték függvényt, amely felett már elvetjük a talált szerkezetet.

Ha az említett módszerek közül egyik sem vezetett eredményre, próbálkozhatunk még a *de novo* szerkezetbecsléssel. Ehhez nem használunk kísérletes adatokat, így önmagában egy nagyon lassú eljárásról van szó, viszont néhány feltételezés lényegesen lecsökkentheti a szóbjövő konformációk terét. Az egymáshoz várhatóan közel elhelyezkedő aminosavak megtalálása ilyen szempontból hasznos lehet, amihez korrelált mutáció analízist végezhetünk. Ennek során többszörös illesztéssel vizsgáljuk a fehérje rokonait, és találhatunk olyan aminosav párokat, amelyek együtt változtak meg a referencia szekvenciához képest. Ilyenkor azzal a feltételezéssel élünk, hogy ha az aminosav mutációja egy másik aminosav mutációjával járt, azok várhatóan egymással kölcsönhatásban álltak, azaz térben közel helyezkedtek el egymáshoz.

A TM fehérjék szerkezetmeghatározásánál külön kihívást jelent, hogy ellentétben más fehérjékkel, itt egy két állapotú (és három fázisú) rendszert kell leírni, ugyanis a fehérje membránba eső régiói lipiddel érintkeznek, az abból kilógó részek viszont vizes közegben helyezkednek el. Erre a körülményre legtöbb módszer nincs felkészítve. Ugyanakkor, alkalmazásukkor a hibák kiszűrése érdekében felhasználhatjuk a TM fehérjék ismert tulajdonságait, például azt, hogy a membrán két különböző oldalán lévő fehérjék sohasem léphetnek kapcsolatba.

3.5. A szerkezeti tér lefedettsége az AlphaFold2 előtt

Korábbi eredmények alapján a humán genom megközelítőleg 25%-a kódol TM fehérjét [5], mégis a megoldott humán TM szerkezetek száma mindössze néhány száz [6]. Ha figyelembe vesszük a homológ szerkezeteket is, akkor ez a szám valamivel magasabb, de ebben az esetben is csak a létező összes TM feltekeredés (fold) alig több mint 60%-a ismeretlen [7].

A korai szerkezetgenomikai projektek egyszerűen nem foglalkoztak a TM fehérjékkel, mivel bonyolult volt a szerkezetük meghatározása kísérletes módszerekkel [8]. A 2010-es évek körül azonban egyre több tanulmány jelent meg a témában. Stratégiákat dolgoztak ki arra, hogy milyen TM fehérje szerkezeteket lenne érdemes megoldani [9], adatbázisok jelentek meg, amelyek tartalmazták a főbb szerkezeti célpontokat [10], és TM fehérje specifikus predikciós módszerek készültek, amelyek képesek megbecsülni, hogy egy fehérje mennyire egyszerűen kristályosítható [11]. Elvben ezek az adatszettek rámutattak a fontos TM fehérjékre, míg a predikciós módszerek segítettek kiválasztani melyik TM fehérje kristályosítása lehetséges. A legújabb megoldások már figyelembe veszik az ortológ fehérjéket is és segítenek kiválasztani azok közül a jól expresszálódó rokonokat, valamint meghatározni a körülményeket, amelyek között a fehérje stabil marad [12].

A kísérletes módszerek mellett a számítógépes módszerek pontossága is elmarad. Homológiamodellezés és foldfelismerés esetén az ismert szerkezetek alacsony száma okoz problémát, míg de novo fehérjemodellezés esetén a TM fehérjék többszörös illesztése okozhat gondot.

Az elmúlt 5 évben azonban a fordulatot nem a szerkezetgenomikai programok által is remélt megoldott röntgen-szerkezetek számának növekedése okozta, hanem a krio-EM, amelynek segítségével nagy mértékben megnövekedett az ismert TM fehérje szerkezetek száma – bár még így sem éri el a kívánt szintet.

A fenti problémákra részben megoldást nyújthat a topológia predikció: ez egy gyors, alacsonyfelbontású szerkezeti leírás, amely tartalmazza a szekvencián belül a TM régiók pontos helyét, a köztes hurkok orientációját a membránhoz képest (organellumot határoló membrán külső vagy belső oldala). Ez a részleges információ már sokféle kísérlet tervezésekor támpontot nyújthat.

3.6. Az AlphaFold2 szerkezetbecslő eljárás

A Critical Assessment of protein Structure Prediction (CASP) egy kétévente megrendezett verseny, ahova bármelyik kutatócsoport jelölheti a szerkezetpredikciós programját. A verseny lényege, hogy addig még nem publikált fehérje szerkezeteken (ennél fogva ezek a szerkezetek nem szerepelhettek egyik eljárás tanítóhalmazában sem) objektíven kiértékelik a versenyző algoritmusokat. A klasszikus, teljesen összekötött neurális hálókat továbbgondoló mély gépi tanulás számos területen hozott már forradalmi sikert. 2020-ban a DeepMind a CASP14 megmérettetésén egy teljesen új szintre emelte a fehérjeszerkezet becslő eljárások pontosságát mély gép tanulás segítségével [13]. A globális távolság teszt (Global Distance Test) egy százaskálán értékeli az algoritmusok pontosságát a kísérletes és a predikált szerkezetek atomi helyzetének összehasonlításával. Az AlphaFold2 (AF2) GDT pontszáma 92,4 volt, míg a második helyezett (ami egyébként az AlphaFold előző verziója) 60 alatti pontszámmal következett.

Az AF2 [14] első lépése nem különbözik a korábban már tárgyalt *de novo* szerkezetbecslő módszerektől: a bemeneti szekvenciához egy rendkívül mély szekvenciaillesztést készít, amelyen korrelált mutáció analízis végezhető. Ezzel egyidőben azonban már egy kiegészítő lépést is elvégez: a saját templát könyvtárában (ahol korábbi kísérletes szerkezetekből származtatott információk vannak) olyan szerkezeteket keres, amelyek szekvenciálisan hasonlóak a lekeresett szekvenciához. Ezek segítségével az AF2 igyekszik megbecsülni az egymáshoz térben közel elhelyezkedő aminosavakat és egy aminosav ‘pár reprezentációt’ készíteni.

Második lépésben ezeket az információkat egy transzformernek továbbítja, amelynek az a feladata, hogy azonosítsa a felhalmozott információkból az igazán fontosakat, azaz amelyek ténylegesen fontosak a szerkezetbecslés szempontjából. Az AF2 transzformer része 48 iteráción keresztül pontosítja és finomítja a többszörös illesztést, valamint a ‘pár reprezentációt’. A harmadik modul a transzformertől kapott információk segítségével készít egy statikus háromdimenziós modellt a fehérjéről.

A fenti eljárás iteratív, a harmadik lépés után ismét a második következik, és a kezdetin felül az AF2 még 3 iteráción keresztül finomítja a szerkezetet. Míg a kezdeti lépésekben a lehetséges hiba magas, és minden egyes lépés után eltérő szerkezetet becsül az AF2, a harmadik iterációban már közel egyforma szerkezeteket predikál.

Az AF2 teljesen új dimenziót nyitott a szerkezeti biológia területén, annak egyik legfontosabb módszerét produkálva az elmúlt évekből. Az AF2-t az eredeti célján felül lehet használni rendezetlen régiók becslésére, variáns effektus analízisre vagy fehérje-fehérje kölcsönhatások vizsgálatára [15]. Az AF2 segítségével mára több százmillió fehérje szerkezetét becsülték meg, így most már az azonosított szekvenciák nagy részéhez létezik egy nagy megbízhatóságú becsült szerkezet [16].

Bár az AF2 pontosságát megmérték egy független halmazon, ezek általában többnyire globuláris fehérjéket tartalmaznak. Az AF2 TM fehérjéken való pontosságáról korábban született cikkek egy-egy fehérjecsaládra vagy típusra koncentráltak. A TmAlphaFold adatbázis az első olyan forrás, amely tartalmazza az AF2-vel becsült összes TM fehérje szerkezetet, valamint a membránsík helyzetét a fehérjéhez képest. A TmAlphaFold adatbázis ezenkívül a szerkezetek értékelését is tartalmazza, geometriai és szerkezeti szempontok alapján – bár ez csak egy intuitív hibaskála, amelynek a részletes értékelése még nem történt meg.

Bár a teljes humán proteomon vizsgálták már hogyan változtatta meg az AF2 a szerkezeti tér lefedettségét [17], de kizárólag TM fehérjéken ezt még senki nem elemezte nagyskáláson. Mivel a TM fehérjék a feltekeredés szempontból egy speciális osztályt képeznek, ezek külön, kritikus vizsgálata egy fontos lépés működésük megértésében és hozzájárul a hatékony gyógyszertervezéshez.

4. Célkitűzés

Az AlphaFold2 (AF2) egészen új megközelítésbe helyezte a szerkezeti biológia területét az elmúlt évben, azonban hatékonyságát transzmembrán (TM) fehérjéken továbbra sem tisztázott. A bizonytalanság oka, hogy az AF2 tanítása során az ismert TM szerkezetek száma igen alacsony volt, így azt várjuk, hogy a nagy számú kísérletesen meg nem oldott humán TM fehérje számos AF2 szerkezeti modellje hibákkal terhelt lehet. Munkám során az alábbi célkitűzésekkel igyekeztem elemezni ezt a problémát:

1. Annak a megállapítása, hogy előre definiált metrikák alapján lehetséges-e hatékonyan elválasztani a jó és alacsony minőségű AF2 szerkezeteket.
2. A humán TM fehérje szerkezeti tér minőségének összehasonlítása az AlphaFold2 megjelenése előtt és után.
3. Annak vizsgálata, hogy az AF2 szerkezetek használhatóak-e homológ fehérjék keresésére, illetve azonosíthatóak-e a segítségükkel már ismert fehérjecsaládok.
4. Olyan fehérjék és fehérjecsaládok meghatározása, amelyekhez továbbra sem tartozik jó minőségű kísérletes vagy számítógépes szerkezet.

5. Adatok és módszerek

5.1. Bioinformatikai módszerek

Módszer neve	Leírás	Elérhetőség
Basic Local Alignment Search Tool (BLAST)	A BLAST [18] biológiai (nukleinsav vagy aminosav) szekvenciák összehasonlításán alapuló kereső program, használatával hasonló szekvenciával rendelkező DNS, RNS vagy fehérjeláncok kereshetők egy háttér adatbázisban.	https://blast.ncbi.nlm.nih.gov/Blast.cgi
CD-HIT	A CD-HIT [19] nukleotid vagy aminosav szekvenciák összevetésére alkalmas program, a betáplált fehérjéket a felhasználó által meghatározott azonossági küszöbértéknek alapján klaszterekbe csoportosítja. Segítségével kiszűrhetők a közelítőleg redundáns szerkezetű fehérjék egy adatszettből.	http://weizhong-lab.ucsd.edu/cd-hit
HHBlits	A HHBlits [20] egy rejtett Markov-modellek páronkénti illesztésén alapuló érzékeny fehérjeszekvencia illesztő program, amellyel távoli homológok kereshetők egy háttéradatbázisban.	https://toolkit.tuebingen.mpg.de/tools/hhblits
Foldseek	A Foldseek [21] egy fehérjeszerkezet-összehasonlító program, ami a hasonló programoknál nagyságrendekkel gyorsabb, azokkal összemérhető pontosságot képes produkálni, így nagy adatszetek esetén is alkalmas homológ fehérjék keresésére.	https://search.foldseek.com/search
TM-Align	A TM-Align [22] egy szekvenciafüggetlen fehérjeszerkezet-összehasonlító algoritmus.	https://zhanggrou.p.org/TM-align
AlphaFold 2 (AF2)	Az AF2 [14] egy mesterséges intelligencián alapuló fehérjeszerkezet predikciós program, amely mélytanulási algoritmusok felhasználásával minden korábbi fehérjestruktúra becslő programnál nagyobb pontossággal képes meghatározni egy fehérje háromdimenziós szerkezetét annak aminosav szekvenciája alapján (ld. 3.6).	https://www.deepmind.com/research/highlighted-research/alphafold
Hierarchikus klaszterezés	A klaszteranalízis során dimenziócsökkentéssel csoportokba (klaszterekbe) rendezhetőek nagyobb adattömbök elemei. A klaszterezést a scikit-learn python programcsomag segítségével végeztem.	https://scikit-learn.org/
Saját scriptek	Munkám során az adatbázisok előkészítési és szűrési lépéseit Python programozási nyelven írtam.	-

1. táblázat. A munkám során felhasznált bioinformatikai módszerek összefoglalása

5.2. Adatbázisok

5.2.1. Széles körben használt adatbázisok

Módszer neve	Leírás	Elérhetőség
Universal Protein Resource (UniProt)	Az UniProt [23] a biológia területén egyik leggyakrabban használt fehérje adatbázis, amelyben megtalálható egyes fehérjék lokalizációja, funkciója és sok egyéb tulajdonsága, más adatbázisokhoz való kereszthivatkozásokkal együtt.	https://www.uniprot.org
Protein Data Bank (PDB)	A PDB [24] adatbázis fehérjék, nukleinsavak és komplexek térszerkezetét tartalmazza, atomi koordináták formájában.	https://www.rcsb.org/
CATH	A CATH [25] egy fehérjeszerkezet-osztályozó adatbázis, ami fehérje domének evolúciós kapcsolatairól nyújt információt, azok kísérletesen meghatározott szerkezetei alapján, automatikus módszerek és kézi annotáció keverékével.	https://www.cathdb.info
Gene Ontology (GO)	A GO [26] egy olyan könyvtár (controlled vocabulary), amely számítógépes műveletekhez is használható modellt állít organizmusok biológiai és kémiai funkcióiról.	http://geneontology.org/
Gyógyszercél-pont fehérjék listája	Az Egyesült Államok Élelmiszer- és Gyógyszerügyi Hivatala (FDA) által jóváhagyott gyógyszercélpontok listáját a Human Protein Atlas weboldaláról töltöttem le.	https://www.proteinatlas.org/search/protein_class%3AFDA+approved+drug+targets

2. táblázat. A munkám során felhasznált adatbázisok összefoglalása

5.2.2. Specifikus adatbázisok

5.2.2.1. PDBTM

A PDBTM [6] (<http://pdbtm.enzim.hu>) a PDB adatbázis TM fehérje szerkezeteinek gyűjteménye, amit a TMDET algoritmus PDB-n való lefuttatásával hoztak létre. A TMDET algoritmus a membrán lokalizációját geometriai megközelítésből igyekszik detektálni, különböző szögben végzett szeletelések mellett a lipidek által hozzáférhető részeket keresve. Az adatbázis fehérjeihez két fontos fájl tartozik, az egyik a pdb fájl, ami a fehérje szerkezetét tartalmazza úgy elforgatva, hogy az x és y tengelyek párhuzamosak legyenek a membránsíkkal. A másik egy xml formátumú fájl, ami tartalmazza a szekvencia egyes régióinak lokalizációját, így a TM régiók pontos helyét is.

5.2.2.2. Orientations of Proteins in Membranes (OPM)

Az OPM [27] (<https://opm.phar.umich.edu>) adatbázis TM fehérjék olyan PDB bejegyzéseinek gyűjteménye, amik a hozzájuk számított membránsík orientációját is tartalmazzák. Itt a membrán lokalizációját a PDBTM-mel ellentétben biofizikai megközelítéssel detektálták. Ez az adatbázis csak a fehérjék pdb fájljait tartalmazza, azokban dummy atomokkal jelölve a membránsíkot, így a szekvenciák pontos TM régióinak meghatározásához saját programokat írtam.

5.2.2.3. Human Transmembrane Proteome (HTP)

A HTP [5] (<http://htp.enzim.hu>) adatbázis humán alfa-helikális TM fehérjék szekvenciáit, és azok predikált topológiáit tartalmazza (azaz a membránrégiók pontos helyét és azok orientációját). Az adatbázis topológia predikcióit a CCTOP [28] segítségével végezték. A CCTOP egy konszenzus módszer, ami 10 másik egyedi módszer eredményét foglalja magában, amelyeket kiegészít kísérletes (topológiát indikáló kísérletek, pl.: foszforiláció, glikoziláció; szerkezeti adatok a PDBTM adatbázisból), valamint bioinformatikai adatokkal (domén információ).

5.2.2.4. AlphaFold Database (AFDB)

Az AFDB [16] (<https://alphafold.ebi.ac.uk>) az AF2 által aminosav szekvenciákból jósolt fehérje szerkezetek adatbázisa. Ez egy viszonylag új szerkezeti adatbázis, ami állandóan frissül. Először csak néhány fontosabb genom került bele, majd a második kiadás a teljes SwissProt adatbázist tartalmazta. A tartalmát előbb közel 1 millió szekvenciára növelték tovább, és a legújabb frissítése már 200 millió fehérje szerkezetet tartalmaz.

5.2.2.5. Transzmembrán AlphaFold adatbázis (TmAlphaFold database)

A TmAlphaFold [29] adatbázis (<https://tmalphafold.ttk.hu>) az AFDB-ből kinyert TM fehérje szerkezetek gyűjteménye. A készítés során az összes AF2 szerkezet fel lett darabolva a PAE mátrixok alapján, majd ezek különböző kombinációin a PDBTM elkészítésénél is alkalmazott TMDET algoritmussal határozták meg a membrán síkját, az esetleges nem kompatibilis részeket kimaszkolva. Erre azért volt szükség, mert az AF2 a kompakt régiókat viszonylag nagy pontossággal képes megbecsülni, azonban az ezek egymáshoz viszonyított orientációját már nem, így egy tipikus hiba lehet, hogy a membrán síkjába belelóg egy globuláris domén.

A TmAlphaFold adatbázis az optimális (sokszor redukált) szerkezet és membránsík meghatározásán túl, a lehetséges geometriai és topológiai hibákat figyelembe véve a szerkezetek minőségét is tartalmazza. A hibák a következők lehetnek:

I) Geometriai hibák

- *Membránsík*: a TMDET nem tudott membránsíkot definiálni (ilyenkor a többi szűrő nem értelmezhető).
- *Szignál peptid*: az érett fehérjéről már levágódó szignál peptid belelóg a membránsíkba.
- *Teljes szerkezet*: A teljes szerkezeten nem lehetett meghatározni a membránsíkot, abból egy vagy több domén ki lett vágva a PAE mátrix alapján.
- *Rövid hélix*: 10 aminosavnál rövidebb alfa-hélix található a membránban.
- *Maszkolt*: Alacsony pLDDT értékkel rendelkező szegmens a membránsíkban.
- *Hiányzó membrán*: A szerkezet szeletelésekor egy fragmensben detektált membránrégió a membránsíkon kívül helyezkedik el a teljes szerkezetben.
- *Domén*: Egy globuláris fragmens belelóg a membránsíkba.

II) Topológiai ellentmondások a CCTOP predikciós programmal való összevetés során

- *Felülbecslés*: A CCTOP predikcióhoz képest extra membránrégió becslése.
- *Alulbecslés*: A CCTOP predikcióhoz képest hiányzik egy vagy több membránrégió a szerkezetből.
- *CCTOP membránsík*: amennyiben nem fed át legalább egy (bitopikus TM fehérjék esetén) vagy két (politopikus TM fehérjék esetén) membránrégió a CCTOP predikció és az AF2 szerkezet között.

Ezek a hibák egy 5 fokozatú skálán vannak összesítve: 0-4: Hibás; 5: Gyenge; 6-8: Elfogadható; 9: Jó; 10: Kiváló.

5.3. A kiértékeléshez használt statisztikai módszerek, metrikák

Az adatok kiértékelése során számos olyan mutatót használtam, amelyek számszerűsítik a felhasznált predikciós módszerek teljesítményét. Ezt úgy tudtam megtenni, hogy a számolt tulajdonságok mellett a létrehoztam teszt adatszeteket is, amelyek azok ellenőrzésére szolgáltak kísérletes és nagy megbízhatóságú adatokkal.

Az adatok kiértékeléséhez konfúziós mátrixokat hoztam létre, amelyek tartalmazzák a helyes és hibás predikciók számát négy kategóriába sorolva: valós pozitív (TP, olyan eset, amikor a predikció és a megfigyelés is pozitív), hibás pozitív (FP, elsőfajú hiba, a predikció pozitív, de a megfigyelés negatív), hibás negatív (FN, másodfajú hiba, a predikció negatív, de a megfigyelés pozitív) és valós negatív (TN, a predikció és a megfigyelés is negatív). Ezekből két fontos érték számítható, amelyek viszonylag pontos képet adnak az osztályozás hatékonyságáról: a Matthews korrelációs koefficiens (MCC), és a balanced accuracy (BACC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$BACC = \frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2}$$

Predikciós módszerek alkalmazásakor eredményként jellemzően egy skálán található értéket kapunk (nagyon sokszor, de nem minden esetben 0-tól 1-ig terjedő számokat), ami a predikció megbízhatóságáról ad információt. Mindig szükség van egy úgynevezett határértékre ('cut value'), ami felett a predikció eredménye pozitív, alatta negatív lesz. Ezt 0 és 1 közötti skálán leggyakrabban 0,5-ként szoktuk meghatározni, de ettől néhány esetben indokolt eltérni. A receiver-operating characteristic (ROC) görbe egy olyan monoton növekvő függvény, amely minden lehetséges határérték esetén ábrázolja valós pozitív ráta (TPR) és hibás pozitív ráta (FPR) értékeket egy grafikonon:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

A ROC görbe alatti terület (AUC) remek mérőszáma a predikciós modellek kiértékelésének, ahol a 0.5 érték egyezik meg a véletlenszerű predikció eredményével.

5.4. Adatszettek

5.4.1. AlphaFold2 humán TM fehérjéken való kiértékelése (humán adatszett)

Az TmAlphaFold adatbázis hatékony elemzéséhez először egy olyan teszt adatszett elkészítése volt a célom, ami megbízható topológiai adatokat tartalmaz humán TM fehérjéről, amivel megfelelően ellenőrizhetők a TmAlphaFold adatbázisban található transzmembrán régió becslések. Ehhez az OPM és PDBTM adatbázisokból egyaránt kigyűjtöttem minden humán TM fehérje szerkezetet, majd kinyertem belőlük a szekvencia, illetve topológia információkat. Az PDBTM adatbázissal ellentétben az OPM adatbázisnál nem állt rögtön rendelkezésemre topológiai információ. Itt a rendelkezésre álló pdb fájlban először megkerestem a megadott membránsík koordináta határait, majd azokat az aminosavakat, amelyek koordinátaik alapján ebbe a tartományba estek, TM régió részeinek vettem, a többit nem, így megkaptam a szekvenciához rendelhető topológiát.

A HTP adatbázisban található szekvenciákra lefuttattam a BLAST algoritmust a PDBTM és az OPM adatbázisok ellen, és listáztam a 40%-os szekvencia azonosságot meghaladó találatokat. Minden illesztett szekvenciára visszatérképeztem az ahhoz tartozó topológiát, majd összevettem az így kapott HTP, OPM és PDBTM topológiákat. Ha az OPM és PDBTM által minden TM régió megfelelően le volt fedve, elfogadtam a megfelelő HTP topológiát.

A fenti találatokat kiegészítettem azokkal a HTP adatbázisban található fehérjékkel, ahol a megbízhatósági érték 95% felett volt. A két adathalmazt ezután összevontam, és a CD-HIT programmal 40% azonosságig szűrtem a szekvenciákat - így biztosítva, hogy redundáns találatoktól mentes adatszettet hozok létre. Az adatszett 759 fehérjét tartalmaz, amelyekből 203-hoz létezik kísérletes szerkezet.

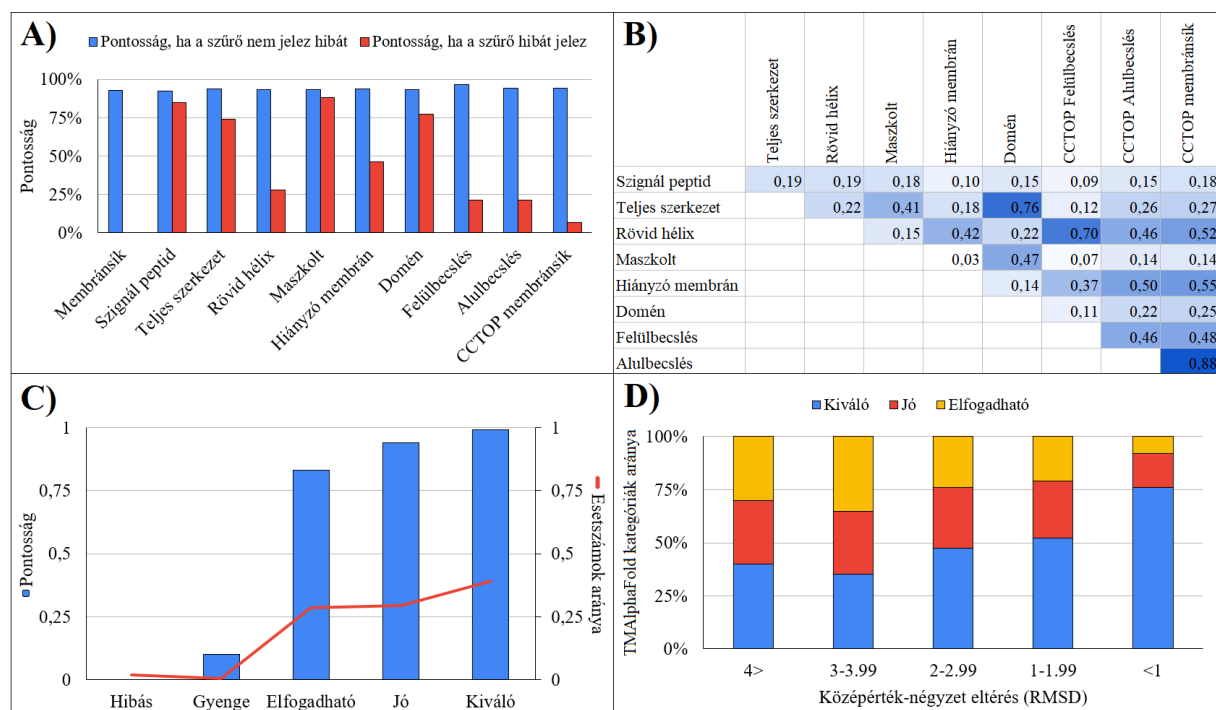
5.4.2. Homológ fehérjék keresése AlphaFold2 szerkezetek segítségével (CATH adatszett)

Annak vizsgálatára, hogy használhatóak-e AF2 szerkezetek homológ fehérjék keresésére, szükségem volt egy olyan adatszetre, ami megbízható, kézzel annotált adatokat tartalmaz fehérjecsaldokról. Erre a célra a CATH adatbázist választottam, amelyből kikerestem azokat a humán TM fehérjékre vonatkozó bejegyzéseket, amelyek a fehérje PDBTM bejegyzése alapján annak összes TM régióját lefedik, és tudtam hozzájuk HTP bejegyzést társítani. Végül az így kapott fehérje listából azokat a fehérjéket ítéltm homológoknak, amelyek a CATH szerint azonos szupercsaládba tartoztak. Az adatszett 290 fehérjét tartalmaz, 33 elkülöníthető osztályba rendezve azokat.

6. Eredmények és kiértékelés

6.1. A TmAlphaFold adatbázis kiértékelése humán transzmembrán fehérjéken

Az AlphaFold2 kiértékelését először egy, a kiértékelés során referenciaként szolgáló ‘humán adatszett’ megalkotásával kezdtem. Ennek lépéseit a dolgozat *Adatok és módszerek* fejezetében már részleteztem. A kiértékelés során a TmAlphaFold adatbázisra támaszkodtam, amely tartalmazza a szerkezetekkel kapcsolatos konfliktusokat.



2. ábra. A TmAlphaFold adatbázis szerkezeteinek értékelése.

A) Helyes topológiák aránya a különböző szűrők ‘helyes’/‘hibás’ jelzései esetén **B)** A TmAlphaFold szűrőinek egymáshoz viszonyított korrelációja. **C)** A topológiák pontossága a TmAlphaFold minőségi kategóriái szerint, valamint az egyes kategóriákba eső fehérjék aránya a ‘humán adatszetten’. **D)** A TmAlphaFold minőségi kategóriák aránya a becsült és kísérletes szerkezetekben lévő atomi koordináták középérték-négyzet eltéréseinek (RMSD) függvényében.

Következő lépésben összevettem a TmAlphaFold adatbázis topológiáit a ‘humán adatszett’ megfelelő fehérjéinek topológiáival, és ahol a vizsgált topológiák megegyeztek, azt helyes szerkezetnek vettem. Úgy gondolkoztam, hogy amennyiben az AF2 szerkezetbecslésből származtatott topológia helyes, az erősen indikálja a szerkezet helyességét is. Ennek a kiértékelésnek az eredményeit összevettem a TmAlphaFold adatbázis saját, geometriai hiba

keresésén, illetve becsült topológiákhoz való hasonlításon alapuló szűrőivel, azaz megvizsgáltam a helyes topológiák arányát amikor egy szűrő azt jelzi, hogy egy fehérje helyes, valamint amikor hibát észlel (2/A ábra). A szűrők változó hatékonysággal működnek: pl a 'szignál peptid' és a 'maszkolt' közel ugyanolyan arányban tartalmaznak helyes fehérjéket 'hibás' és 'helyes' jelölés esetén (ennek fő oka, hogy ezek a részek a membránsík detekció során ideiglenesen törlésre kerülnek, de ettől még hibaként vannak számontartva). A geometriai szűrők közül a 'rövid hélix' és a 'hiányzó membrán' esetben működik a legjobban, míg a topológia szűrők közül a CCTOP membránsík - ezekben az esetekben, ha a szűrő nem jelez hibát a pontosság magas, míg hiba esetén alacsony.

Ezután a szűrők esetleges redundanciájának vizsgálatára megnéztem azok egymáshoz való korrelációját is. Erre azért van szükség, hogy láthatóvá váljon, hogy az egyes szűrők nem redundánsak és a különböző hibatípusok más fehérjéknél kerülnek elő. Az egyes szűrők általában nem mutatnak magas korrelációt (2/B ábra), a geometriai szűrők közül a 'teljes szerkezet' és a 'domén' esetén látható magasabb érték (0,76) - ennek magyarázata, hogy amennyiben egy domént el kellett távolítani a teljes szerkezetből, az valószínűleg azért történt mert zavarta a membránsíkot. Korreláció figyelhető meg továbbá az 'alulbecslés' és a 'CCTOP membránsík' szűrőknél, amelyek eltérő interpretációi a becsült topológiával való összevetésnek.

Azt is megvizsgáltam, hogy amennyiben a pontszámokat összesítjük a TmAlphaFold adatbázisban ajánlott rendszer szerint, akkor az egyes minőségi kategóriákban mekkora a pontosság és mekkora a fehérjék aránya (2/C ábra). Jól látszik, hogy a magasabb minőségi kategória nagyobb pontossággal jár (kivételt jelent a 'gyenge' kategória, de ez betudható a nagyon alacsony mintaszámnak (5 fehérje)) és a tesztalmazból ide több fehérje tartozik.

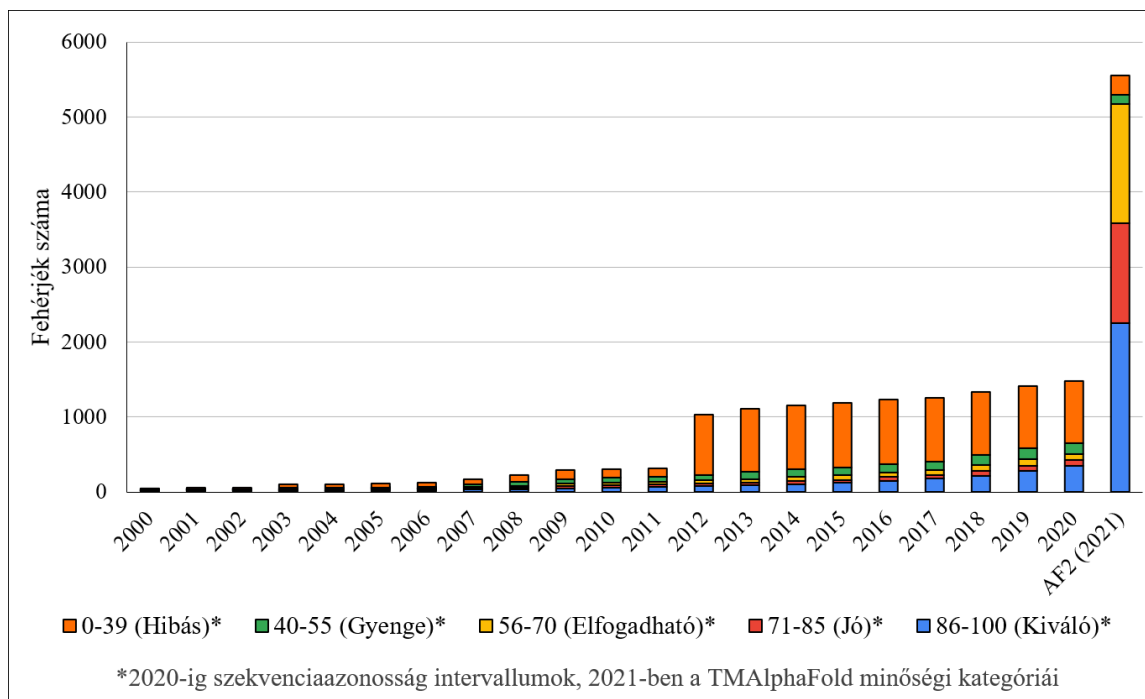
Ezenkívül vizsgáltam még a TmAlphaFold szerkezeteket topológiai helyett szerkezeti megközelítésből is úgy, hogy összevettem a TmAlphaFold becsült szerkezeteinek és kísérletesen megoldott szerkezetek atomi koordinátáit a TM-Align programmal, és ezek távolságának középérték-négyzet eltérését vettem (Root Mean Square Deviation: RMSD). Ennél az összehasonlításnál a 'humán adatszett' csak egy részét tudtam használni, ahol volt megoldott kísérletes szerkezet (203 fehérje). Az eredmények alapján minél nagyobb az RMSD értéke, annál több fehérje kerül a gyengébb kategóriákba (ha az $RMSD > 4$, a kiváló fehérjék aránya csak 40%), míg alacsony RMSD (< 1) esetén a kiváló kategória dominál, ahova a fehérjék 80%-a került (2/D ábra).

A TmAlphaFold ilyen, szerkezeti megközelítésből való vizsgálatának van pár fontos hátránya: I) az általam felhasznált teszhalmaz nem lehet teljesen független, ugyanis az AF2-t az összes 2020-ig bezárólag megjelent szerkezetek felhasználásával tanították. Fontos megemlíteni, hogy a mágneses magrezonanciából (Nuclear Magnetic Resonance: NMR) származó szerkezeteket nem vették figyelembe a tanítás során. A megoldott TM fehérjék legtöbbször röntgenkrisztallográfia vagy krio-elektronmikroszkópia (krio-EM) eredményei. Viszonylag kevés az NMR segítségével meghatározott TM fehérje szerkezet, ugyanis ehhez túl nagyok a fehérjék, pedig az ilyen szerkezetek egyébként alkalmasak lennének független teszhalmaz létrehozására. Felmerül, hogy a 2021 utáni szerkezetek, vagy az NMR-rel megoldott szerkezetek alkalmasak lehetnek független teszhalmaz létrehozására, azonban reálisan ezeknek a szerkezeteknek a nagy része homológ a korábban megoldott szerkezetekkel. II) A gyenge és hibás kategóriába nem kerültek fehérjék a teljes szerkezetek összehasonlításakor: ennek éppen az az oka, hogy a fehérjék, amikhez rendelkezésre áll kísérletesen megoldott szerkezet, tipikusan benne voltak az AF2 tanító halmazában, így ezeken nagyobb annak pontossága.

6.2. A humán TM szerkezeti tér lefedettségének minősége az AF2 előtt és után

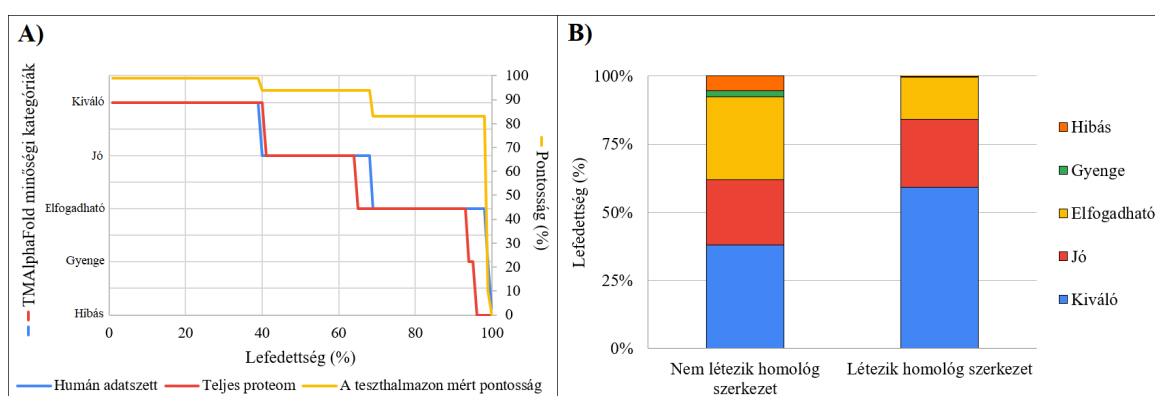
A következőkben a AF2 megjelenésének a megbízható TM fehérje szerkezetek számára gyakorolt hatását igyekeztem felderíteni. Ennek vizsgálatához első lépésben megvizsgáltam, hogy az egyes humán TM fehérjékhez a különböző években mi volt a (szekvenciális azonosságot figyelembe véve) legközelebbi szerkezet, majd összegeztem minden évre, hogy hány fehérjéhez létezik különböző küszöbértékek mellett ismert szerkezet. Ezeket az értékeket a kísérletes meghatározáshoz köthető évszám függvényében oszlopdiagramon ábrázoltam (2020-ig, 3. ábra), majd ehhez hozzávettem a TmAlphaFold összes szerkezetének minőségi kategóriák szerinti eloszlását (2021-es év, 3. ábra).

Jól látszik, hogy az AF2 által számolt megbízható („Kiváló”) szerkezeti modellek száma sokszorososa a kísérletes úton előállított szerkezetek számának (még úgy is, hogy nem használunk szűrőt a szekvencia azonosságra). Az AF2 tehát hatalmas előrelépést jelent a megbízható TM fehérje szerkezeti modellek számában. A diagramon egy nagy ugrás látható 2011 és 2012 között a megoldott szerkezetek számában, ez annak tulajdonítható, hogy 2012-ben meghatározták egy humán szaglőreceptor fehérje szerkezetét - mivel ez az egyik legnagyobb fehérjecsaládba tartozik, így ugrásszerűen megnőtt azoknak a fehérjéknek a száma, amikhez létezik közeli szerkezet.



3. ábra. A teljes humán TM proteom szerkezeti lefedettsége az AF2 előtt (kísérletes szerkezetekhez való szekvenciális azonosság) és becsült szerkezetek minősége az AF2 megjelenése után (TmAlphaFold minőségi kategóriák).

A továbbiakban azt is vizsgáltam, hogy a 'humán adatszett' képest hogyan alakul az AF2 lefedettsége a teljes TM proteomon, a TmAlphaFold adatbázis minőségi kategóriáit felhasználva (4/A ábra): Mind a tesztalmazon, mind a teljes proteomon az adatok ~40%-a esik a kiváló kategóriába, ami a tesztalmazon mérve 99%-os pontosságot jelez. A fehérjék 65-70%-a esik a jó kategóriába, ahol még mindig viszonylag magas a pontosság (93%).



4. ábra. A teljes proteom vizsgálata a TmAlphaFold minőségi kategóriával.

A) A TmAlphaFold minőségi kategóriák arányainak összevetése a 'humán adatszett' és a teljes proteom fehérjéin, valamint a tesztalmazon mért pontosság. **B)** A TmAlphaFold minőségi kategóriák arányainak összevetése a teljes proteomon, homológ szerkezettel rendelkező, illetve nem rendelkező fehérjéin.

Ezenkívül a teljes proteomon is összevettem a TmAlphaFold adatbázis minőségi kategóriáinak arányát olyan fehérjéknél, amelyekhez nem rendelhető közeli PDB szerkezet, szemben azokkal, amelyekhez lehet homológ szerkezetet találni (4/B ábra). Itt látványosan jobb minőségű szerkezetek figyelhetők meg, amennyiben van közeli kísérletesen megoldott szerkezet. Ennek az lehet az oka, hogy az AF2 kissé túlilleszt, tehát azokon a fehérjéken működik igazán jól, amelyeken tanítva lett. Ez nem meglepő, gépi tanulási módszereknél szinte elkerülhetetlen valamilyen szintű túlillesztés a tanítóhalmazra.

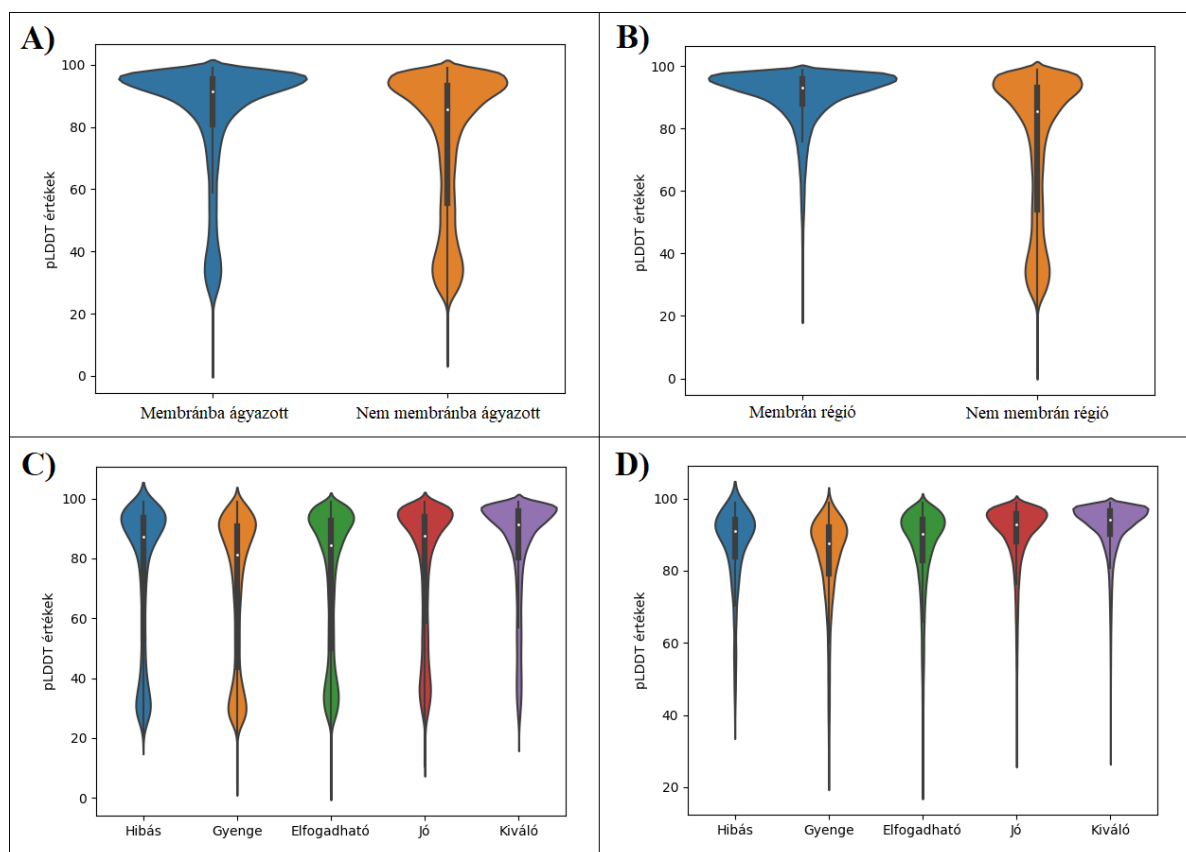
Az előzőekben a TmAlphaFold kiértékelése során az egyes fehérjék helyességét sok esetben a topológia segítségével határoztam meg. A következő lépésben annak az összevetése volt a célom, hogy a különböző TmAlphaFold besorolások és HTP megbízhatósági értékek esetén milyen arányban egyeznek a szekvenciából becsült, valamint az AF2 szerkezeteken a TMDET segítségével meghatározott membránsíkokból származtatott topológiák. A HTP megbízhatósági értékeket 5 csoportba osztottam. Az eredmények alapján amennyiben a HTP megbízhatósági értéke 90% felett van, valamint a TmAlphaFold besorolása kiváló, a topológiák 98,63%-ban egyeznek (3. táblázat). Az is látszik, hogy ha csak a független változókat vesszük figyelembe (HTP megbízhatóság vagy TmAlphaFold besorolás), akkor a magasabb kategória magasabb egyezést mutat. Ez a vizsgálat arra mutat rá, hogy amennyiben két fundamentálisan eltérő predikció (HTP, AF2 + TmAlphaFold) mindegyike magabiztos eredményt ad, az eredmény nagyobb eséllyel lesz megbízható.

		HTP megbízhatóság					
		<60	60-69	70-79	80-89	90+	
TMAIphaFold besorolás	Kiváló	0,95	0,89	0,92	0,98	0,99	0,97
	Jó	0,74	0,55	0,52	0,74	0,93	0,75
	Elfogadható	0,34	0,34	0,39	0,57	0,81	0,55
	Gyenge	0	0	0	0,31	0	0,07
	Hibás	0	0	0	0	0	0
		0,46	0,42	0,55	0,81	0,89	

3. táblázat. A TmAlphaFold és HTP topológiák összevetése különböző megbízhatósági szinteken (vasatag: legalább 50 adat).

Kíváncsi voltam még az AF2 által használt aminosavankénti megbízhatósági becslés értékek (pLDDT) eloszlására is az egyes TmAlphaFold szerkezetekben. Először a membránba ágyazott és nem membránba ágyazott aminosavak pLDDT értékei közötti különbséget vizsgáltam meg, amihez vettem az összes TmAlphaFold bejegyzésben definiált membránsík Z koordináta határait, és a megfelelő pdb fájlból kigyűjtöttem az összes olyan aminosavhoz tartozó pLDDT

értéket, amelynek Z koordinátája a membránsík határai közé esett, illetve külön azokat, amik ezen kívül esetek (5/A ábra). Hasonló módon, a TmAlphaFold adatbázisban definiált topológiai információk alapján a membránrégióba, és a nem oda eső aminosavak pLDDT értékeit is összegyűjtöttem (5/B ábra).



5. ábra. Violin plotok a pLDDT értékek eloszlásáról különböző TmAlphaFold szerkezetekben

A) A membránba ágyazott, és nem membránba ágyazott aminosavak pLDDT értékeinek eloszlása az összes TmAlphaFold szerkezetben. **B)** A membránrégióban elhelyezkedő, és a nem membránrégióban elhelyezkedő aminosavak pLDDT értékeinek eloszlása az összes TmAlphaFold szerkezetben. **C)** Az aminosavak pLDDT értékeinek eloszlása az összes TmAlphaFold szerkezetben, TmAlphaFold minőségi kategóriák szerint. **D)** A membránrégióban elhelyezkedő aminosavak pLDDT értékeinek eloszlása az összes TmAlphaFold szerkezetben, TmAlphaFold minőségi kategóriák szerint.

Megfigyelhető, hogy a membránba ágyazott, illetve a membránrégióba eső aminosavak esetén jóval nagyobb arányban vannak a magas pLDDT értékek, mint a többi esetben. Ez igazolja azt, hogy a membránban található aminosavak predikciója valamivel biztosabb. A membránon kívül megfigyelhető alacsonyabb pLDDT értékek egyébként vélhetően gyakran rendezetlen

régiókra vonatkoznak [30], hiszen az alacsony pLDDT értékek és a rendezetlenség több megfigyelés alapján is korrelálnak egymással. Az is megfigyelhető, hogy a membránba ágyazott esetben (ahol a feltétel nem kötötte ki, hogy az adott rész egy tényleges membránrégió legyen) előfordulnak alacsony megbízhatósági aminosavak is, szemben a kizárólag membránrégiókkal. Ezekben az esetekben az AF2 sokszor egy flexibilis régiót húzott keresztül a membránon - aminek az oka, hogy a tanítás során nem volt kikötve, hogy a kettős lipidrétegen áthaladó résznek periodikus másodlagos szerkezetet kell felvennie.

TmAlphaFold minőségi kategóriák alapján válogatva is ábrázoltam az ide tartozó fehérjék aminosavjainak pLDDT értékeit (5/C ábra). Itt az elvárásaimnak megfelelően jobb minőségi kategóriákhoz átlagosan magasabb pLDDT értékek is tartoztak. Ezt az ábrát úgy is elkészítettem, hogy csak a membránrégióba eső aminosavak pLDDT értékeit választottam ki (5/D ábra), ebben az esetben is megfigyelhető ez a trend, illetve az alacsonyabb pLDDT értékek hiányából itt is látszik, hogy a membránban található aminosavak predikciója biztosabb. Az egyetlen kivételt a 'Hibás' kategória jelenti, ahol az átlagos pLDDT érték magasabb, mint a soron következő, gyenge kategóriában - ilyen esetben pl. a membrándomén kompakt módon meg lett jósolva, de egy hélix a membránon kívül lett feltekerve.

6.3. Szekvencia és szerkezet alapú módszerek összehasonlítása homológ fehérjék keresésére

Az AF2 kiértékelését egy másik nézőpontból folytattam, megvizsgáltam, hogy mennyire hatékonyan lehet AF2 szerkezetek segítségével homológ (pontosabban paralóg) fehérjéket keresni. Viszonyítási alapnak a hasonló, szekvencia alapú homológ keresést használtam. A homológ keresési eljárások hatékonyságának ellenőrzésére hoztam létre a 'CATH adatszettet', benne fehérjék megbízható familiáris kapcsolataival, ennek lépéseit a dolgozat *Adatok és módszerek* fejezetében már részleteztem.

A szekvencia alapú keresést a HHBlits algoritmussal végeztem. Ehhez HTP adatbázis minden fehérje szekvenciáján lefuttattam a HHBlits-et a teljes adatbázis ellen. A futtatások eredményeképpen minden fehérjéhez egy fájlt kaptam, benne az összes illesztéssel, és az illesztések e-value értékeivel. Visszatérképeztem az illesztett fehérjékre azok topológiáját a HTP alapján, majd csak azokat az illesztéseket tartottam meg, ahol az illesztett rész mindkét fehérje összes TM régióját lefedte. Az így kapott fehérjepárokat vettem össze a 'CATH adatszett' fehérjepárjaival, a predikció eredményeként az adott fehérjepárokhoz tartozó e-value

értéket felhasználva (amennyiben a HHBlits mindkét irányból megtalálta a fehérjepárt, a kisebb e-value értékkel, azaz a biztosabb találattal számolva).

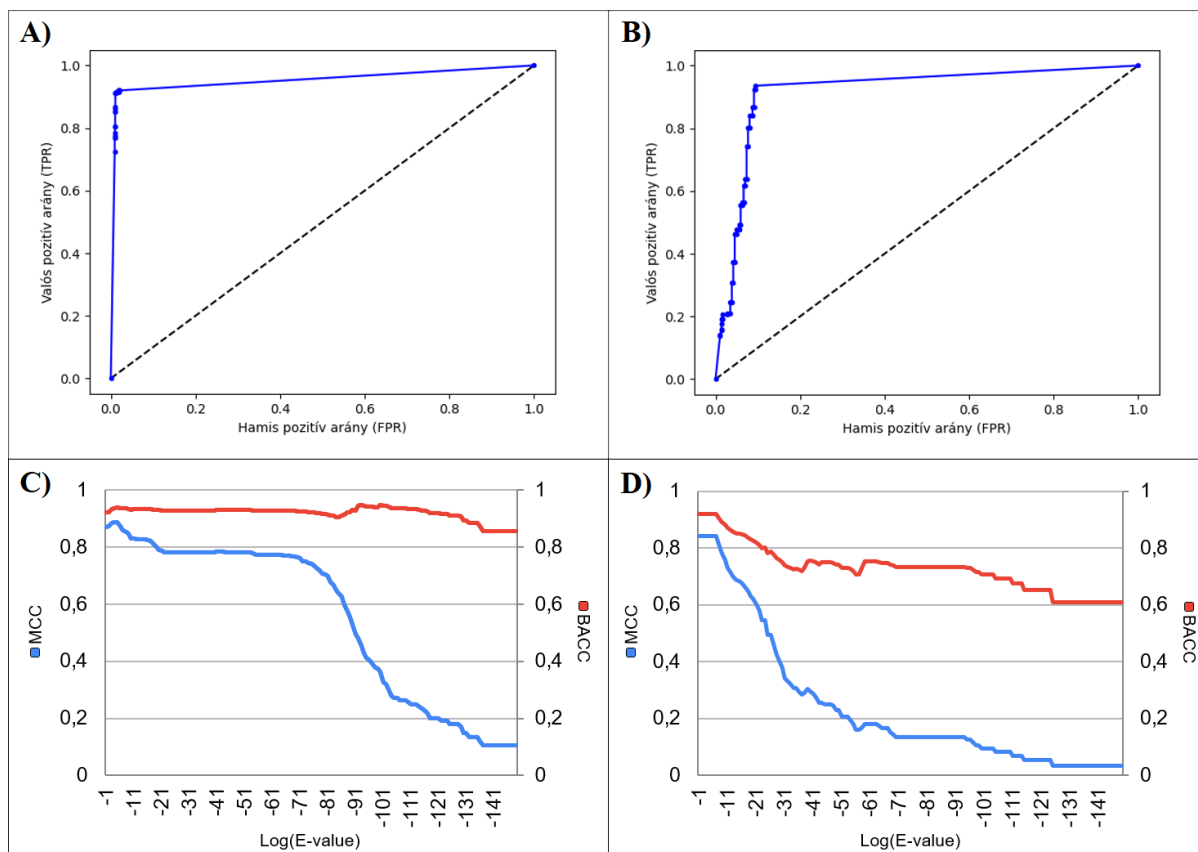
Módszer	Feltétel	Legmagasabb MCC
HHBlits	-	0,895
HHBlits	numTM	0,892
HHBlits	TopFilter	0,728
Foldseek	-	0,874
Foldseek	numTM	0,838
TM-Align	-	0,809

4. táblázat. A szekvencia és szerkezet alapú homológ fehérje kereső algoritmusok eredményeinek legmagasabb MCC értékei különböző szűrők mellett. Vastaggal jelöltem a későbbi összehasonlításokban használt algoritmust.

A kiértékelés során különböző szűrőket próbáltam ki a megfelelő kapcsolatok kiválasztásakor (4. táblázat), így tehát az összes találatot elfogadó adatszetten túl készítettem egy olyan adatszettet, amiben csak az azonos számú TM régióval rendelkező fehérjék kapcsolatát fogadtam el (NumTM), illetve egy olyat is, amelyben ezen túl az illesztésekre is kikötéseket tettem, különböző topológiai megfontolások alapján (minden membrán régió legalább 5 aminosav hosszúságban átfedjen, illetve a TM régióba beszúrt gap-ek száma ne legyen több mint 5) (TopFilter), ezzel is igyekezve javítani a pontosságot.

Fontos kiemelni, hogy mindkét esetben (HHBlits és Foldseek) végül alkalmaztam egy olyan szűrőt, amely ellenőrzi a TM régiók számát - ennek az az oka, hogy bár a pontosság nagyon kicsit alacsonyabb, de a futási idő nagyságrendekkel gyorsabb ilyen esetben, ami a későbbi proteom szintű lekeresésnél volt fontos szempont.

A predikció teljesítményét vevő működési karakterisztika (Receiver Operating Characteristic: ROC) görbén ábrázoltam (6/A ábra), ez alapján egy meglehetősen jó predikcióról van szó, a ROC görbe alatti terület (Area Under Curve: AUC) értéke 0,954. Ábrázoltam még a különböző e-value küszöbértékek mellett számolt Matthews korrelációs koefficiens (Matthew's Correlation Coefficient: MCC) és pontosság (Balanced Accuracy: BACC) értékeket is (6/C ábra) az ideális küszöbérték kiválasztása érdekében. Látszik, hogy a maximum pont a görbe elején található, tehát csak nagyon kevés, magas e-value-val rendelkező fehérjepárt nem érdemes a találatok közé venni. Amennyiben a HHBlits egy kapcsolatot megtalált, az legtöbbször helyes.

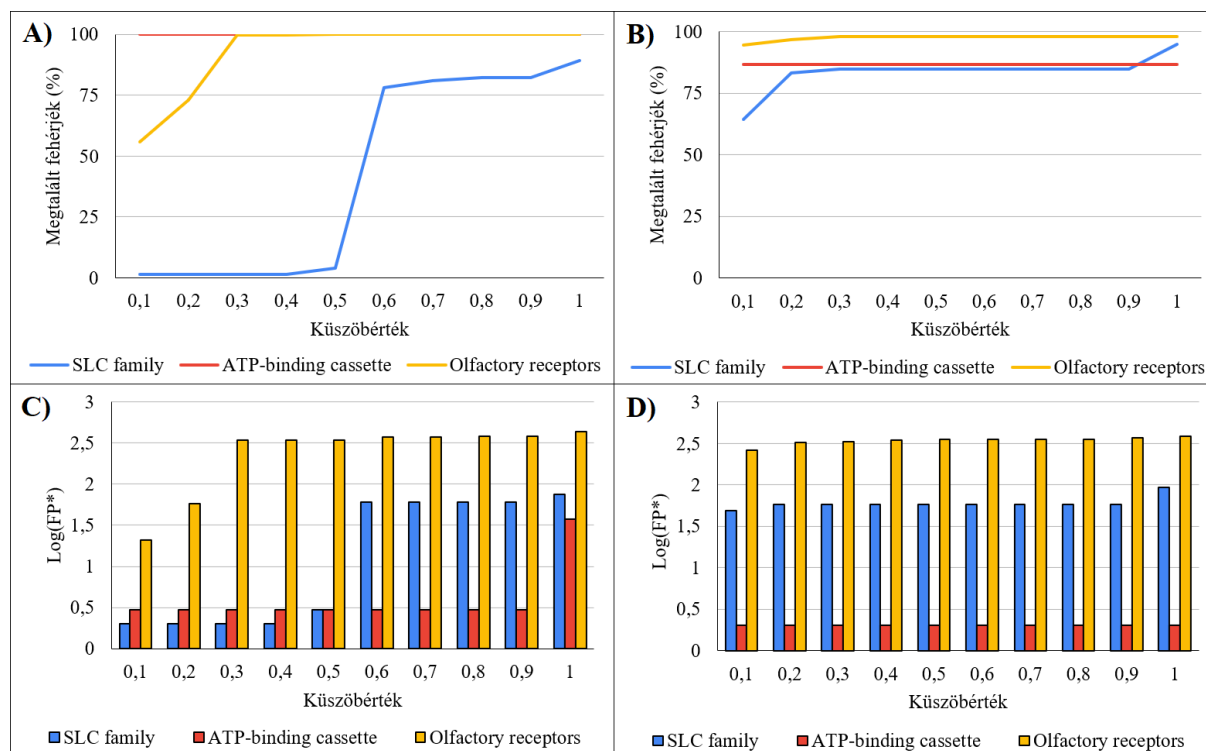


6. ábra. A szekvencia és szerkezet alapú homológ fehérje kereső algoritmusok teljesítmény mutatói.

A) A HHBlits (numTM) eredményeinek ROC görbéje. **B)** A Foldseek (numTM) eredményeinek ROC görbéje. **C)** A HHBlits eredményeinek MCC és BACC értékei az e-value függvényében. **D)** A Foldseek eredményeinek MCC és BACC értékei az e-value függvényében.

Az AF2 szerkezetek alapján történő homológ keresésre a Foldseek programot választottam, a TmAlphaFold adatbázis fehérje szerkezeteit felhasználva. A TM-Align program használatával is megpróbálkoztam a célra, és ez a tesztek alapján nagyon kevéssel pontosabb eredményeket is adott a Foldseeknél, azonban ennek futása jóval erőforrásigényesebb és lassabb, ezért végül a Foldseek használata mellett döntöttem. A Foldseek a talált fehérje kapcsolatokat egy fájlba írta ki a hozzájuk tartozó e-value értékekkel, ezekre a kapcsolatokra ezúttal is készítettem egy szűrőt, amely kiválasztja az azonos számú TM régióval rendelkező fehérjéket. Ebben az esetben is vizsgáltam a pontosságot szűrővel és anélkül is. A 'CATH adatszettel' összehasonlítási alapként véve ismét számoltam ROC görbét (6/B ábra), amely ezúttal is pontos és specifikus predikcióra utal. Itt az AUC értéke valamivel elmarad a szekvencia alapú keresésétől, de így is meglehetősen magas (0,918). A különböző e-value küszöbértékek mellett számolt MCC és BACC értékeket ábrázolva (6/D ábra) látszik, hogy a maximum pont itt is a görbe legelején

van, tehát nagyjából az e-value értékétől függetlenül, amiket a Foldseek kapcsolatként megtalált, azok jellemzően valóban homológ fehérjék voltak.



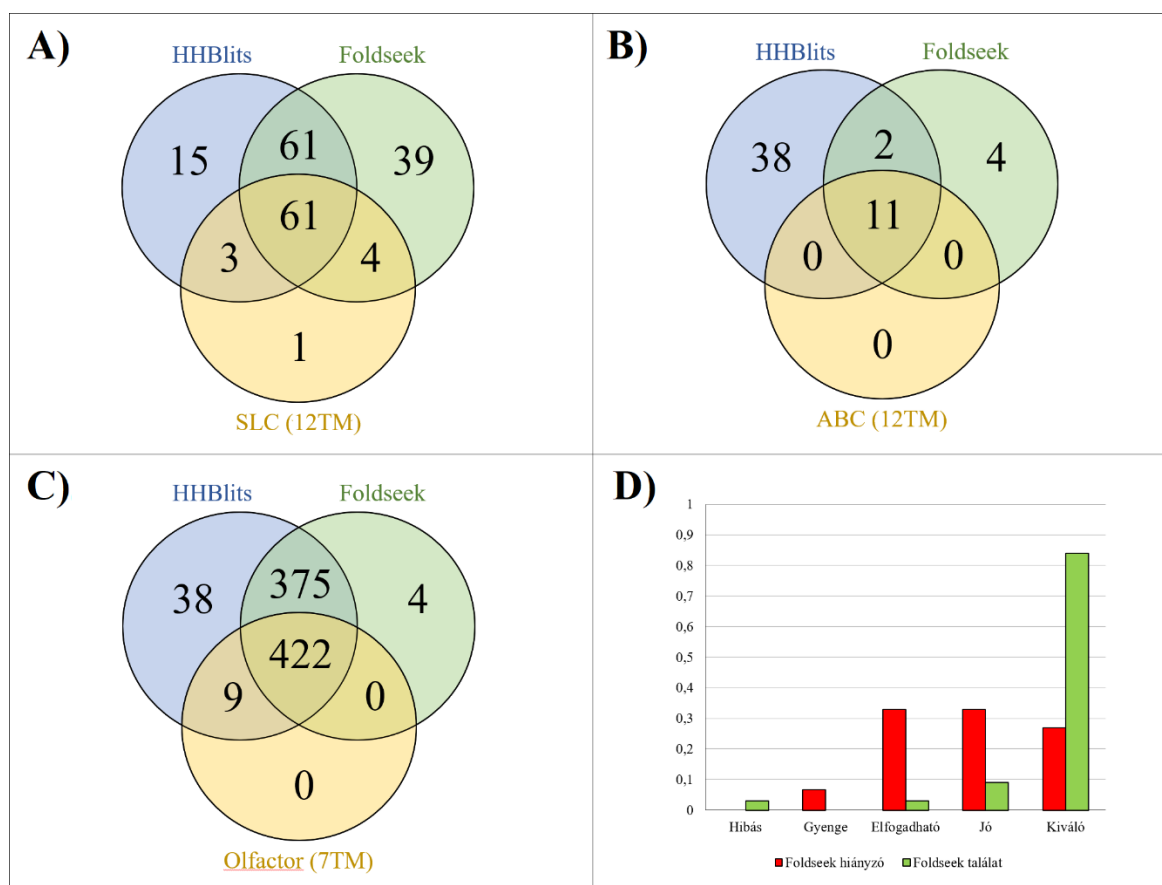
7. ábra. A hierarchikus klaszterezés eredményeinek kiértékelése.

A) A vizsgált fehérjecsaládok közül megtalált fehérjék aránya eltérő küszöbértékek mellett a HHblits eredményei alapján. **B)** A vizsgált fehérjecsaládok közül megtalált fehérjék aránya eltérő küszöbértékek mellett a Foldseek eredményei alapján. **C)** A hamis pozitív* (lsd. szöveg) találatok aránya eltérő küszöbértékek mellett a HHblits eredményei alapján. **D)** A hamis pozitív* (lsd. szöveg) találatok aránya eltérő küszöbértékek mellett a Foldseek eredményei alapján.

Következő lépésben a legoptimálisabb lekeresési módokat (lsd. 3. táblázat, vastaggal) kiterjesztettem a teljes humán TM proteomra. Arra voltam kíváncsi, hogy a HHblits és Foldseek által megtalált kapcsolatok alapján lehet-e jól elkülönülő fehérje csoportokat (klasztereket) létrehozni. Erre a célra összevonó (agglomerative) hierarchikus klaszterezés segítségével a fehérjéket csoportokra osztottam az eredmények alapján. A klaszterezés során állítható a távolsági küszöbérték, ami megadja, hogy az algoritmus milyen megengedően vonja össze az elemeket nagyobb csoportokba. Három fontosabb fehérjecsaládot választottam (SLC fehérjecsalád, ATP-binding cassette (ABC) transzporter fehérjék és szaglás receptorok (Olfactory receptors)), amikhez viszonyíthattam a talált fehérje csoportokat. A csoportokat aszerint határoztam meg, hogy a fehérje teljes neve, vagy a génnev tartalmazta-e a megfelelő

kulcsszavakat. Ezután elvégeztem a klaszterezést 0,1-től 1,0-ig minden küszöbérték mellett. Az ideális küszöbérték megtalálása érdekében ábrázoltam fehérjecsaládokból megtalált fehérjék arányát a különböző küszöbértékek mellett (7/A-B ábra). Megfigyelhető, hogy 0,6-os küszöbérték felett már minden esetben megtalálható a fehérjecsaládok fehérjéinek nagy része.

Emellett a valóban helyes fehérjecsaládok megtalálásának van még egy fontos mérőszáma: azoknak a fehérjéknek a száma, amiket egy csoportba helyez az algoritmus, pedig nem biztos hogy oda tartoznak. Ezeket az egyszerűség kedvéért a továbbiakban hamis pozitív*-nak (FP*) jelöltem, bár igaz, hogy ezek nagy részéről valójában nem tudhatjuk biztosan, hogy hamis pozitívak, hiszen tartozhatnak valóban az adott fehérjecsaládba, csak erről a gén- vagy fehérjenévalapján nincsen információnk. Ezeknek a FP* fehérjéknek a számát ábrázoltam a különböző küszöbértékek mellett (7/C-D ábra). A HHBlits esetén látható egy kiugrás 0,6-os küszöbértéknél, pont ahol a helyes találatok aránya is emelkedik. A Foldseek esetén itt az látszik, hogy kevésbé számít a küszöbérték, a FP fehérjék száma állandó.



8. ábra. A klaszterezéssel készített csoportok összehasonlítása referencia fehérjecsaládokkal.

A) SLC fehérjecsalád, **B)** ABC fehérjecsalád, **C)** szaglőreceptorok, illetve a **D)** Foldseek futtatásban szereplő, és a hiányzó fehérjék szerkezeti minőségének összehasonlítása.

Venn-diagramokon foglaltam össze a fehérjék eloszlását a három kiválasztott fehérjecsalád, illetve a HHBlits és Foldseek eredményeiből klaszterezéssel (engedékeny, 1,0-ás küszöbértéket kiválasztva) létrehozott fehérjecsoportok között (8/A-C ábra). Itt az látszik, hogy a HHBlits lényegesen több találatot eredményez az ABC transzporter és szaglóreceptor kategóriában, de ezek között feltehetően több a hibás, míg a Foldseekkel egy specifikusabb klaszterezés hozható létre.

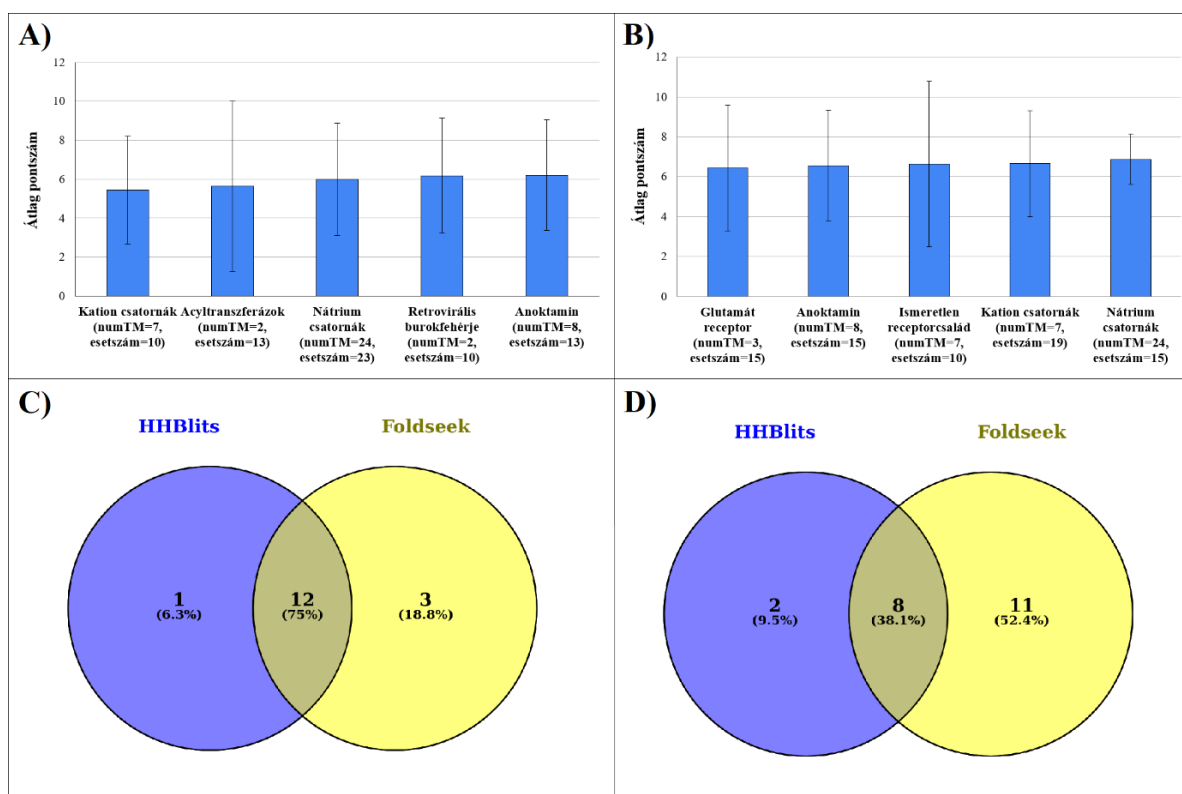
Végül arra voltam kíváncsi, hogy a Foldseek jellemzően milyen minőségű AF2 szerkezetekhez tudott, illetve nem tudott homológokat találni. Ehhez a TMAAlphaFold minőségi kategóriáit használtam fel, és ezeknek arányát ábrázoltam a Foldseek találattal rendelkező (True positive: TP) és nem rendelkező (False negative: FN) fehérjék esetén (8/D ábra). Szembetűnő, hogy a Foldseek a jó minőségű szerkezetek esetén képes megtalálni a rokon fehérjéket, azonban, ha a TmAlphaFold minőségi kategória rosszabb, a Foldseek is nagyobb eséllyel véti el a találatot.

Eredményeim alapján az AF2 alapú szerkezeti lekeresés hatékony távoli homológ fehérjék megtalálására TM fehérjék esetén, és hozhat olyan eredményeket, amelyek szekvenciális kereséssel nem találhatóak meg - azonban a lekeresés erősen függ az AF2 becsült szerkezetének minőségétől. Továbbá arra jutottam, hogy a klaszterek létrehozásakor meglehetősen nehéz egyértelmű határokat húzni a fehérjecsoportok között, nem egyszerű olyan e-value értéket, topológiai vagy egyéb szabályt találni, amivel univerzálisan szét lehetne választani a különböző fehérjecsaládokat. Ennek fényében úgy döntöttem, hogy a továbbiakban nem a klaszterezéssel megállapított családokat használom majd, hanem minden fehérjére külön definiálom a vele egy családban lévő fehérjéket, a homológia keresések eredményei alapján. Így ezek a kapcsolatok gyakran nem kölcsönösek, nem feleltethetőek meg egymásnak kölcsönösen egyértelműen.

6.4. Gyengébb minőségű szerkezetbecslések elemzése szerkezeti, funkcionális és gyógyszertervezési szempontból

A folytatásban a HHBlits és a Foldseek segítségével létrehozott fehérjecsoportokból összeállítottam két úgynevezett 'most wanted' listát, amikben azok a fehérjecsaládok találhatóak, melyekben átlagosan a legalacsonyabb a TmAlphaFold által definiált szerkezeti minőséget jellemző pontszám (tehát ebben az esetben nem az 5 minőségi kategóriát használtam, hanem az azokat definiáló pontszámot). Korábban minden egyes fehérjéhez meghatároztam milyen más fehérjével van kapcsolatban. Először a legalacsonyabb átlagos pontszámú fehérjecsoportokat egy úgynevezett mohó ('greedy') algoritmussal gyűjtöttem össze, kizárólag a minimum 10 taggal rendelkező csoportokat számolva. Röviden, ez egy iteratív eljárás: miután

a legalacsonyabb átlag pontszámú csoportot kiválasztottam, az itt szereplő fehérjéket a továbbiakban nem vettem figyelembe, hanem töröltem a fehérjelistából és a rokonok közül is, így minden lépésben újradefiniálva a rokonok halmazát - ezzel elkerültem, hogy ugyanaz a fehérje kétszer is a listában szerepelhessen. Ezt az iterációt ötször végeztem el. A bitopikus 1 TM fehérje csoportokat nem vettem bele az listába, mert azokon tipikusan bizonytalanabbak a predikciós eljárások, így könnyebben téves eredményre vezethetnek. Ezek után a kapott fehérjecsoportok fehérjéire rákerestem az UniProt adatbázisban, és intuitívan kerestem a számukra egy nevet, ami a legtöbb fehérjét jellemzi a csoportban. Az így kapott fehérjecsaládok átlagos pontszámát ábrázoltam a hozzájuk számolt szórásokkal együtt mind a HHBlits, mind a Foldseek eredményeiből létrehozott fehérjecsaládok esetén (9/A-B ábra). Ezek közül főleg azok a csoportok érdekesek, ahol alacsony a szórás, mert ezeknél nem áll rendelkezésünkre egy magas pontszámú szerkezet sem.

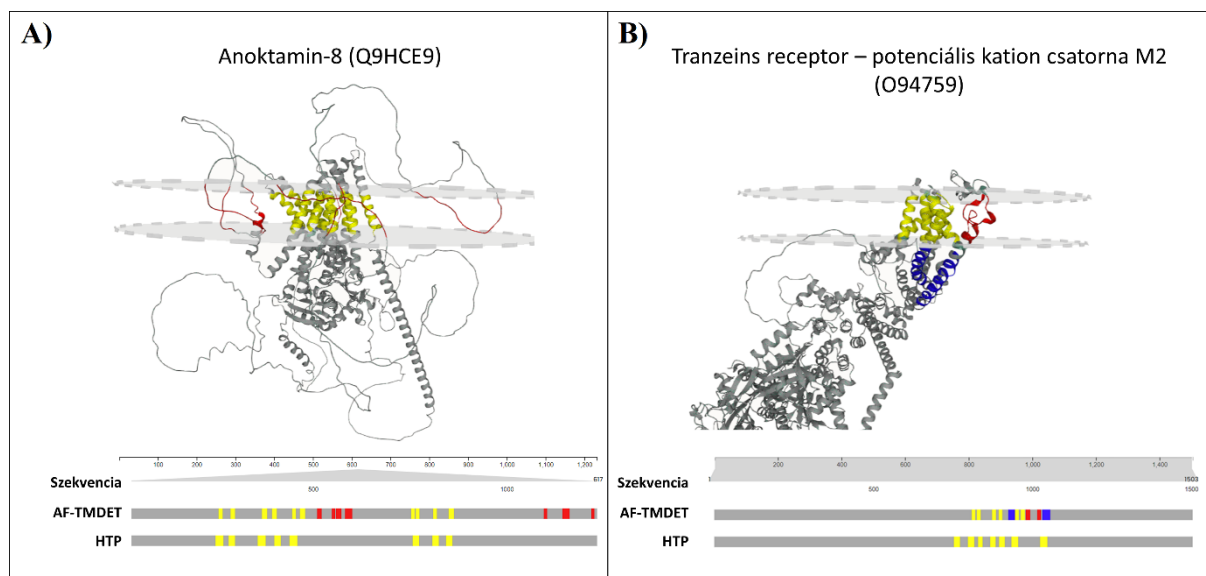


9. ábra: A HHBlits és a Foldseek alapú megközelítés 'most wanted' listái.

A) A HHBlits eredményeiből létrehozott fehérjecsaládok 'most wanted' listájának átlagos TmAlphaFold szerkezeti minőség pontszámai **B)** Foldseek eredményeiből létrehozott fehérjecsaládok 'most wanted' listájának átlagos TmAlphaFold szerkezeti minőség pontszámai **C)** A HHBlits és Foldseek által megtalált anoktamin fehérjék összehasonlítása Venn-diagramon **D)** A HHBlits és Foldseek által megtalált kation csatorna fehérjék összehasonlítása Venn-diagramon

Látszik, hogy három fehérjecsalád mindkét módszer 'most wanted' listájára felkerült: az anoktaminok, a kation csatornák, illetve a nátrium csatornák. Ezek közül az anoktaminok és a kation csatornák esetén Venn-diagramon ábrázoltam a HHBlits és a Foldseek alapú megközelítéssel megtalált fehérjék összevetését (9/C-D ábra). Míg az anoktamin (anion csatorna) fehérjék esetén nagy az átfedés, a kation csatornák esetében a Foldseek több találatot eredményez. A nátrium csatornák esetén nem készítettem ilyen összehasonlítást, mert a HTP hibásan 23 TM régiót becsül a család fehérjeinek egy részéhez, így a HHBlits-en alkalmazott 'numTM' szűrő itt hibás összerendeléseket is eredményezett.

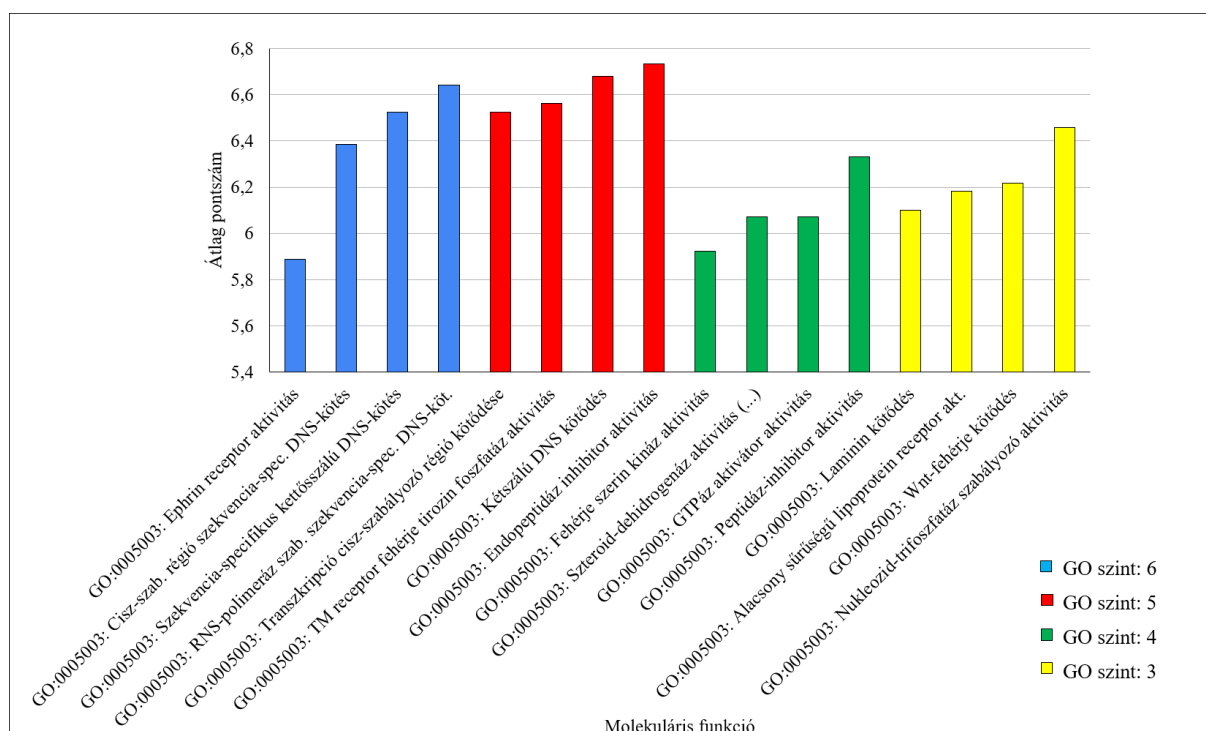
Egy-egy példafehérjén vizsgálva láthatóvá válnak az AF2 hibái a fenti fehérjecsaládok esetében. Az anoktamin esetén a membránba belelógna flexibilis hurkok, valamint részlegesen egy alfa-hélix is megjelenik a membránban. A topológiabecsléssel összevetve extra membránrégiók figyelhetők meg, valamint egyes membránrégiók el vannak tolva (10/A ábra). A potenciális kation csatornák esetén a TM domén nem megfelelően van becsülve: TM régiók hiányoznak, máshol felül vannak becsülve, valamint a membránba eső részek másodlagos szerkezete hibás (10/B ábra).



10. ábra: Hibás AlphaFold2 szerkezetek a TmAlphaFold adatbázisból.

Felül a szerkezet látható a membránsíkkal együtt. Pirossal vannak jelölve a hibásan membránba kerülő részek, kékkel azok a szegmensek, amiknek a kettős lipidrétegben kéne elhelyezkedniük, mégis azon kívül esnek. Alul az TmAlphaFold adatbázisban található topológia és a CCTOP becslés összevetése látható, hasonló színezéssel. A) Anoktamin-8 B) Potenciális kation csatorna.

Azt is megvizsgáltam, hogy a GO könyvtárban melyek azok a molekuláris funkciók, amelyekhez átlagosan alacsony TmAlphaFold szerkezeti minőség pontszámok tartoznak. Ehhez meghatároztam az összes GO molekuláris funkció szintjét a fán, majd mindegyikhez összegyűjtöttem, hogy a TmAlphaFold adatbázis fehérjéi közül hányhoz rendelhető hozzá. A 3. és 6. szint között jellemzően már nem túl általános, de nem is túl specifikus kifejezések találhatók, így ezekről a szintekről választottam ki a négy legalacsonyabb átlagos TmAlphaFold minőség pontszámmal rendelkező kifejezéseket, amelyekhez legalább 10 fehérje rendelhető, ezeket ábrázoltam (11. ábra).



11. ábra: A legalacsonyabb átlagos TmAlphaFold szerkezeti minőség pontszámmal rendelkező GO molekuláris funkciók különböző szinteken.

A talált molekuláris funkciók között láthatók a specifikus, és kevésbé specifikus csoportok. A specifikusabb (magasabb GO szintű) csoportok között többször felbukkan DNS kötés funkció, aminek a nukleáris folyamatoknál van szerepe. Ez azért érdekes, mivel ez arra utalhat, hogy a nukleuszba lokalizálódó fehérjék modellezése során kevésbé működik jól az AF2, amit a sejtmag membránjának a plazmembrántól valamennyire eltérő felépítése indokolhat. A kevésbé specifikus funkciók közül a Wnt-fehérje kötődés jelenléte érdekes az alacsony pontszámú csoportok között, hiszen ez egy fontos jelátviteli útvonal, ami számos fejlődési és regulációs folyamatban részt vesz, és rákos megbetegedéseknél is központi szerepe van, ezért fontos lehet ezen fehérjéről való szerkezeti információink pontosítása.

Végül a Human Protein Atlas gyógyszercélpont fehérje listája alapján összegyűjtöttem azokat a gyenge vagy hibás TmAlphaFold minőségi kategóriájú szerkezettel rendelkező TM fehérjéket, amelyek az FDA által jóváhagyott gyógyszercélpontok. Itt egyéni annotációval további információkat kerestem a talált fehérjékről, ezeket az adatokat egy táblázatban foglaltam össze (5. táblázat).

Uniprot AC	Gén leírás	TM régiók száma	TmAlphaFold kategória	PDB lefedettség	Megjegyzés
P16109	Szelektin	1	Hibás	22.00%	Csak a globuláris rész lefedett
P29317	Ephrin receptor	1	Hibás	95.00%	Teljes szerkezet lefedett (NMR)
Q14534	Epoxidáz	bizonytalan	Hibás	80.00%	Csak a globuláris rész lefedett
P14060	Dehidrogenáz	2	Gyenge	-	Nincs szerkezet
P16234	Növekedési faktor receptor	1	Gyenge	86.00%	Globuláris rész lefedett, homológ fehérje 3D szerkezete ismert
P26439	Dehidrogenáz	bizonytalan	Gyenge	-	Nincs szerkezet
P51172	Nátrium csatorna	2	Gyenge	-	Nincs szerkezet
Q00975	Kalcium csatorna	24	Gyenge	100%	Teljes szerkezet lefedett (Cryo-EM, AF tanítás után)

5. táblázat: Gyenge és hibás minőségű AF2 szerkezettel rendelkező gyógyszercélpont TM fehérjék és az azokhoz társítható PDB szerkezetek tulajdonságai.

Látható, hogy ezek között vannak fehérjék, amelyekhez létezik teljes lefedett szerkezet, amely tartalmazza a membránrégiót is. Ilyen a *P29317* fehérje, amelyhez NMR-rel, vagy a *Q00975*, amelyhez krio-EM-el meghatározott szerkezet található meg. Ezek a fehérjék viszont nem szerepeltek az AF2 tanítóhalmazában, az első esetben azért, mert a szerkezet NMR-rel lett megoldva, míg a második esetben a PDB adatbázisba csak a tanítás után került be a bejegyzés. Ezek a fehérjék modellezési szempontból kevésbé érdekesek, egyéb algoritmusokkal viszonylag könnyen megoldható a szerkezetük. A *P16234* fehérjének ugyan csak globuláris része van lefedett a PDB-ben, de létezik olyan viszonylag közeli homológ fehérjéje, aminek teljes szerkezete ismert, így vélhetően ez is könnyedén modellezhető. Az igazán érdekes találatok a *P14060*, *P26439* és *P51172* fehérjék, amikhez nem áll rendelkezésünkre kísérletes szerkezet (homológ sem), de gyenge minőségi besorolást kaptak. A *P16109* fehérje is fontos találat, ugyanis ennek a fehérjének is csak egy globuláris doménje megoldott szerkezetű, a többi része hiányzik a röntgenből, beleértve a TM régiót. A jövőben ezeknek a fehérjeszerkezeteknek a megoldása potenciálisan segítheti új gyógyszerhatóanyag kifejlesztését.

7. Összefoglalás

Az elmúlt évek során a szerkezeti biológia egyik legfontosabb eredménye lett az AlphaFold2 (AF2) szerkezetbecslő program megjelenése. Az AF2 új szintre emelte az összes atom pontos helyzetét meghatározó szerkezetbecslések pontosságát, és megreformálni látszik mind az egyedi fehérje, mind a nagyskálás vizsgálatok területét. Bár az AF2 átlagos pontossága vitathatatlanul magas, a transzmembrán (TM) fehérjék speciális tulajdonságaik miatt nehezen értékelhetőek hagyományos eszközökkel.

Munkám során létrehoztam egy olyan kiértékelési rendszert, ami képes meghatározni, hogy proteom szinten milyen hatékony az AF2 algoritmus a TM fehérjéken. Az eredmények alapján az AF2 ezeken a fehérjéken is nagy pontossággal működik, azonban egyszerű geometriai és topológiai megközelítésekkel megtalálhatóak azok az esetek, ahol feltehetően hibás eredményt ad. Miután egy jól ellenőrzött tesztalmazon megbizonyosodtam az AF2 szerkezeteinek objektív kiértékelhetőségéről, az analízist kiterjesztettem a teljes humán TM proteomra is. Az eredmények alapján az AF2 a teljes humán proteom 40%-án nagyon pontos (99%), és további ~25%-on is erős eredményt ért el (93%). Az eredményeim segíthetnek abban, hogy ezeket a fehérjéket pontosan azonosítsuk a létrehozott metrikák segítségével.

A szerkezeti biológia egyik fontos lépése, hogy meghatározzuk azokat a fehérjéket, amelyek egymáshoz hasonlóak, hiszen ennek segítségével számos egyszerűsítéssel élhetünk vizsgálatok során. Ez szinte minden bioinformatika analízis vagy program egyik első lépése. Az AF2 ebben is új lehetőséget teremt, hiszen mostantól nem csak a szekvencia információt használhatjuk keresésre, hanem a becsült szerkezeteket is. Bemutattam, hogy az AF2 segítségével becsült szerkezet alapú keresés hogyan viszonyul a klasszikus szekvencia alapú keresésekhez TM fehérjék esetén, és hogy ennek segítségével olyan távoli homológokat is megtalálhatunk, amikre korábban nem volt lehetőség csak a szekvencia figyelembevételével.

A fenti megközelítések kombinálásával a dolgozatomban választ kerestem arra is, hogy mik azok a fehérjecsaládok, amelyeknek a szerkezeti reprezentációja mindig alacsony minőségű. Ezeket a halmazokat szekvenciális, szerkezeti és funkcionális szempontból is vizsgáltam. Az eredmények alapján az AF2 gyengébb eredményt ér el bizonyos kation csatornákon és az anoktamin (anion csatorna) fehérjéken, valamint a nukleáris lokalizációjú TM fehérjéken. Bemutattam azokat a gyógyszercélpontokat, ahol a szerkezet minősége gyenge és nem létezik közeli kísérletesen megoldott szerkezet. Ezeknek a fehérjéknek a szerkezetmegoldása kiemelt fontosságú lehet a jövőben.

8. Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Dr. Dobson Lászlónak, amiért fáradhatatlan iránymutatásával és támogatásával nagyban hozzájárult, hogy ez a dolgozat elkészülhessen.

Köszönöm csoportvezetőmnek, Dr. Tusnádý Gábornak, hogy lehetővé tette, hogy a csoportjában dolgozzam, illetve Buday Lászlónak és Hunyady Lászlónak amiért az intézetükben végezhettem munkámat.

Természetesen hálával tartozom még családomnak, és minden barátomnak, akik az egyetemi éveim során rengeteget támogattak.

9. Hivatkozások

- [1] Anfinsen, Christian B. "Principles that govern the folding of protein chains." *Science* 181.4096 (1973): 223-230.
- [2] Levinthal, Cyrus. "Are there pathways for protein folding?." *Journal de chimie physique* 65 (1968): 44-45.
- [3] Dill, Ken A. "Polymer principles and protein folding." *Protein Science* 8.6 (1999): 1166-1180.
- [4] Uhlén, Mathias, et al. "Tissue-based map of the human proteome." *Science* 347.6220 (2015): 1260419.
- [5] Dobson, László, István Reményi, and Gábor E. Tusnady. "The human transmembrane proteome." *Biology direct* 10.1 (2015): 1-18.
- [6] Kozma, Daniel, Istvan Simon, and Gabor E. Tusnady. "PDBTM: Protein Data Bank of transmembrane proteins after 8 years." *Nucleic acids research* 41.D1 (2012): D524-D529.
- [7] Oberai, Amit, et al. "A limited universe of membrane protein families and folds." *Protein science* 15.7 (2006): 1723-1734.
- [8] Büssow, Konrad, et al. "Structural genomics of human proteins—target selection and generation of a public catalogue of expression clones." *Microbial cell factories* 4.1 (2005): 1-13.
- [9] Punta, Marco, et al. "Structural genomics target selection for the New York consortium on membrane protein structure." *Journal of structural and functional genomics* 10.4 (2009): 255-268.
- [10] Varga, Julia, et al. "TSTMP: target selection for structural genomics of human transmembrane proteins." *Nucleic acids research* (2016): gkw939.
- [11] Varga, Julia K., and Gábor E. Tusnady. "TMCrys: predict propensity of success for transmembrane protein crystallization." *Bioinformatics* 34.18 (2018): 3126-3130.
- [12] Loukeris, Michael, et al. "Fine Sampling of Sequence Space for Membrane Protein Structural Biology." *Journal of Molecular Biology* 433.15 (2021): 167055.
- [13] Kryzhtafovych, Andriy, et al. "Critical assessment of methods of protein structure prediction (CASP)—Round XIV." *Proteins: Structure, Function, and Bioinformatics* 89.12 (2021): 1607-1617.

- [14] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.
- [15] Akdel, Mehmet, et al. "A Structural Biology Community Assessment of AlphaFold2 Applications." *Nature Structural & Molecular Biology* 29, no. 11 (2022): 1056–67.
- [16] Varadi, Mihaly, et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models." *Nucleic acids research* 50.D1 (2022): D439-D444.
- [17] Porta-Pardo, Eduard, et al. "The structural coverage of the human proteome before and after AlphaFold." *PLoS computational biology* 18.1 (2022): e1009818.
- [18] Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.
- [19] Fu, Limin, et al. "CD-HIT: accelerated for clustering the next-generation sequencing data." *Bioinformatics* 28.23 (2012): 3150-3152.
- [20] Remmert, Michael, et al. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." *Nature methods* 9.2 (2012): 173-175.
- [21] van Kempen, Michel, et al. "Foldseek: fast and accurate protein structure search." *bioRxiv* (2022).
- [22] Zhang, Yang, and Jeffrey Skolnick. "TM-align: a protein structure alignment algorithm based on the TM-score." *Nucleic acids research* 33.7 (2005): 2302-2309.
- [23] UniProt: the universal protein knowledgebase in 2021. *Nucleic acids research*, 2021, 49.D1: D480-D489.
- [24] Burley, Stephen K., et al. "RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences." *Nucleic acids research* 49.D1 (2021): D437-D451.
- [25] Sillitoe, Ian, et al. "CATH: increased structural coverage of functional space." *Nucleic acids research* 49.D1 (2021): D266-D273
- [26] Gene Ontology Consortium. "The gene ontology resource: 20 years and still GOing strong." *Nucleic acids research* 47.D1 (2019): D330-D338.

- [27] Lomize, Mikhail A., et al. "OPM: orientations of proteins in membranes database." *Bioinformatics* 22.5 (2006): 623-625.
- [28] Dobson, László, István Reményi, and Gábor E. Tusnady. "CCTOP: A Consensus Constrained TOPology Prediction Web Server." *Nucleic Acids Research* 43, no. W1 (2015): W408–12.
- [29] Dobson, László, et al. "TmAlphaFold Database: Membrane Localization and Evaluation of AlphaFold2 Predicted Alpha-Helical Transmembrane Protein Structures." *Nucleic Acids Research* 51, no. D1 (2022): D517–22.
- [30] Ruff, Kiersten M., and Rohit V. Pappu. "AlphaFold and implications for intrinsically disordered proteins." *Journal of Molecular Biology* 433.20 (2021): 167208.